www.phmsociety.org

ISBN – 978-1-936263-40-0

PHME 2024

Proceedings of the 8th European Conference of the Prognostics and Health Management Society 2024

Prague, Czech Republic
July 3rd - July 5th , 2024

ISBN – 978-1-936263-40-0

Edited by:
Phuc Do
Cordelia Ezhilarasu

phmsociety

Management team

General operational functions:

Konstantinos Gryllias – KU Leuven – General Chair
Ian Jennions - Cranfield University - General Co-Chair
Octavian Niculita – Glasgow Caledonian University – Finance Chair
Jeff Bird – TECnos – PHMSociety Board of Directors, Sponsorship Chair
Claude Foubert – VERT COM – General Support Chair
Cordelia Ezhilarasu - SLB/Cranfield University – TPC Chair
Phuc Do – University of Lorraine – Proceedings Chair

Specific sessions:

Andrew Starr - Cranfield University - Panels Chair
Bin Zhang - University of South Carolina – Panel Co-Chair
Steve King - Cranfield University - Tutorials Co-Chair
Tingting Zhu - Oxford University - Tutorials Co-Chair
Yvonne Lu - Oxford University - Tutorials Co-Chair
Mihaela Mitici – Utrecht University – Doctorial Symposium Chair
Khanh Nguyen – Tarbes National School of Engineering, Doctorial Symposium Co-Chair

Technical Program Committee members:

Ian Jennions - Cranfield University, UK
Fakhre Ali - Cranfield University, UK
Piero Baraldi – Politecnico di Milano, Italy
Christophe Berenguer -Univ. Grenoble Alpes, France
Nima Yousefi - Lucid Motors, Canada
Pierre Dersin - Lulea University, Sweden
Phuc Do - University of Lorraine, France
Cordelia Ezhilarasu - SLB/Cranfield University, UK
Fink Olga - EPFL, Switzerland
Mihaela Mitici - Utrecht University, Netherlands
Bruce Stephen - University of Strathclyde, UK
Steve King - Cranfield University, UK
Francesco Cadini - Politecnico di Milano, Italy
Ahmed Mosallam – SLB, France
Octavian Niculita - Glasgow Caledonian University, UK
Khanh Nguyen - Tarbes National School of Engineering, France
Zeina Al Masry – Ecole Nationale Superieure, France
Slawomir Nowaczyk - Halmstad University, Sweden
Benoit Iung - University of Lorraine, France
Alexandre Voisin - University of Lorraine, France
Dong Wang - Shanghai Jiao Tong University, China
Ferhat Tamssaouet - University of Perpignan, France

Published by **PHM Society**
Publisher Address:
241 Woodland Drive, State College, PA 16803
http://www.phmsociety.org/journal/publisher

Table of contents

Posters

Dotoral Symposium

A Comparative Study of Semi-Supervised Anomaly Detection Methods for Machine Fault Detection

Dhiraj Neupane, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal

School of Information Technology, Deakin University, Waurn Ponds, VIC 3216, Australia
{*d.neupane, reda.bouadjenek, richard.dazeley, sunil.aryal*}*@deakin.edu.au*

Abstract

Industrial automation has extended machines' runtime, thereby raising breakdown risks. Machine breakdowns not only have economic and productivity consequences, but they can also be fatal. Thus, the early detection of fault signs is essential for the safe and uninterrupted operation of machinery and its maintenance. In the last few years, machine learning has been widely used in machine condition monitoring. Most existing approaches rely on supervised learning techniques, which face challenges in real-world scenarios due to the lack of enough labelled fault data. Additionally, models trained on historical fault data might struggle to detect new and unseen faults accurately in the future. Therefore, this research uses semi-supervised Anomaly Detection (AD) techniques to detect abnormal patterns in machines' vibration signals. As semi-supervised techniques are trained on normal data only, they do not require faulty samples and abnormal patterns are detected based on their deviations from the learned normal pattern. We compared the effectiveness of seven state-of-the-art AD methods, ranging from traditional approaches such as isolation forest and local outlier factor to more recent Deep Learning (DL) approaches based on autoencoders. We evaluated the effectiveness of different feature types extracted from the raw vibration signals, including simple statistical features like kurtosis, mean, peak-to-peak, and more complex representations like the scalogram images. Our study on three public datasets, with unique challenges, shows that the traditional methods based on simple statistical analysis have shown comparable and sometimes superior performance to more complex DL approaches. The use of traditional approaches offers simplicity and lower computational needs. Thus, our study recommends that future researchers start with the traditional approaches first and then jump to DL methods if necessary.

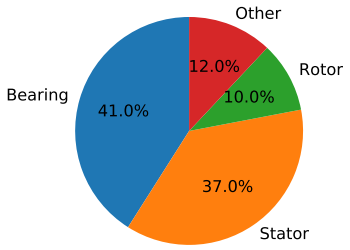


Figure 1. Proportion of machine components failure

1. Introduction

Rotating machinery is a fundamental component of modern industry and has a wide range of applications in practical engineering, including electric machines, trains, turbines, aero-engines, and so on (Jiao, Zhao, Lin, & Liang, 2019). The ubiquitous presence of these devices, from simple mechanical systems to complex nuclear power plants, reflects their critical role in modern industrial processes (Zhong, Zhang, & Ban, 2023). With the advancement of technology and productive growth in modern industry, there has been an increased reliance on machinery, making them frequently operated under adverse and challenging conditions and increased risks of failures. If unattended timely and accurately, these failures can have significant consequences, including decreased production efficiency, financial losses, and, in extreme cases, the potential loss of human lives (Neupane & Seok, 2020). Common failures in electric motors include bearings, stators, rotors, and gearboxes. Figure 1 shows the failure rates of these machinery components. These components are vital for efficient power transmission and operation of machinery. However, continuous use can result in wear, cracks, and defects of these components that can lead to machine breakdowns. Therefore, prompt and accurate fault detection and diagnosis are essential. Thus, timely maintenance of these components is critical to the machine's safe and reliable operation.

Fault diagnosis and maintenance are crucial for improving production efficiency and reducing accident rates in mechanical systems. Both the academic and industrial communities

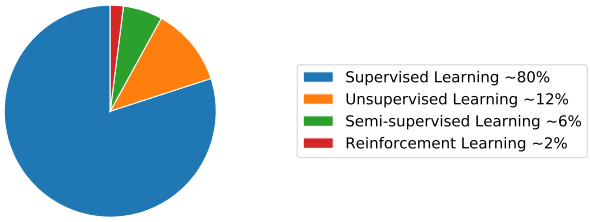


Figure 2. ML techniques used for MFD

have acknowledged the significance of Machinery Fault Diagnosis (MFD), leading to the development of various diagnostic methods for practical applications (Li, Zhang, Qin, & Estupinan, 2020). MFD has become an essential aspect of industrial development and engineering research, and numerous strategies have been developed by researchers, scientists, and engineers through years of innovative and diligent work.

Over the last decade, Machine Learning (ML) techniques have been widely used in MFD. A vast majority (over 80%, see figure 2) of those MFD methods have used supervised learning (SL) approaches (Das, Das, & Birant, 2023) to classify fault types. While such methods can detect faults previously seen, they are unable to detect new or unseen types of faults. Because many modern machines are operated in complex industrial environments, new types of faults can emerge over time. Also, to train a decent model to classify different types of faults, we need a sufficient amount of labelled data for each fault type. The scarcity of labelled data is a challenging problem in real-world industrial settings. Data labelling is an expensive and time-consuming process as it requires domain expertise to manually annotate different types of faults. Moreover, labelled data might not cover the entire spectrum of possible faults, leading to a lack of diversity in the training dataset and potentially limiting the model's ability to generalize to unseen faults.

To show the aforementioned limitations of SL in MFD, we evaluated the capability of the Decision Tree classifier using deep features from the pre-trained ResNet (ResNet-DT) (He, Zhang, Ren, & Sun, 2016) in detecting previously unseen faults. We trained the ResNet-DT model for binary classification (faulty vs. normal type) by excluding certain fault types from the training set, while including all fault types in the test set. The objective is to distinguish between normal operation and any fault condition, rather than identifying specific types of faults. We used 10 runs of a random 70-30 train-test split for each combination of omitting $i = \{0, 1, 2\}$ fault types from the training set. Our results, shown in figure 3, for the Case Western Reserve University (CWRU) datasets show that the ResNet-DT model's performance declines significantly when it encounters fault types that were not present during the training. In the x-axis of figure 3, labels C0, C1, and C2 represent the number of fault types intentionally omitted during the

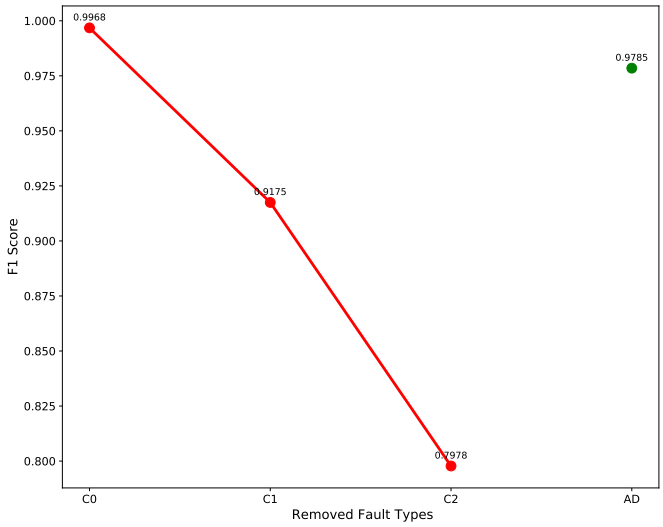


Figure 3. Average F1 score of the ResNet-DT classifier from scalogram images of vibration signals on the CWRU dataset.

model's training phase. C0 indicates that the model is trained with all fault types included. C1 represents the model being trained with one fault type excluded; this is done sequentially for each fault type (first excluding fault type 1, then including it while excluding fault type 2, and so on). Similarly, C2 denotes the exclusion of two fault types simultaneously. The y-axis shows the average F1-score for the classification of the fault condition, corresponding to the different combinations of omissions. Due to the numerous possible combinations of omitted fault types, we calculated and presented the average F1 score. The red dots in the figure denote the respective average F1 score for each fault type omission. In contrast, the green dot represents the F1 score of the Isolation Forest (iF) based semi-supervised Anomaly Detection (AD) method using the same ResNet deep features (ResNet-iF) trained on half of the normal dataset. The other half is concatenated with all fault types together. It is evident from figure 3 that the ResNet-DT model encounters challenges in detecting unknown faults. The trend shows a significant decrease in the F1-score as more fault types are excluded from the training set, underscoring the model's limitations in recognizing unseen machinery faults. In contrast, the ResNet-iF's average F1 score shows the effectiveness of AD methods in detecting unseen faults. The iF, trained on half the amount of the normal state machinery signals and tested on all the fault types along with the other remaining half of the normal data, performed nearly equal ($\tilde{1}.8\%$ lesser) to the ResNet-DT model (trained with 70% data as training) with no classes omitted in training.

Taking the supervised model's ineffectiveness in detecting unseen faults in real-world scenarios as the motivation for this project, we have explored the potential of semi-supervised learning (SSL) based AD algorithms that are trained on normal data only and aim to detect unseen fault types. These

algorithms model the profile of normal vibration signals to distinguish faulty (or abnormal) vibration signals from normal signals. In the real-world scenario, where the availability of normal/healthy machinery data is abundant, these algorithms are very useful and can detect anomalies or faults more easily and quickly than the SL classification models.

The use of SSL in MFD is relatively unexplored. Prior studies employing SSL techniques mostly focus on classifying the faults only. A recent study (Zong et al., 2022) on bearing fault diagnosis of CWRU and Xi'an Jiaotong University dataset focused on the use of SSL. The study utilized a short-time Fourier transform as a preprocessing step and employed SSL with domain adversarial neural network for fault classification and achieved an average accuracy of 96.77%. Another study by Zhang et al. (Zhang, Ye, Wang, & Habetler, 2021) also focused on SSL employing VAE for the classification of bearing faults for the CWRU and University of Cincinnati Intelligent Maintenance System dataset. With 16.67% of labelled data in each class, the accuracy of $\tilde{9}8\%$ was achieved. Moreover, a research (Zhang, Ye, Wang, & Habetler, 2020) addressed bearing AD challenges via few-shot learning based on model-agnostic meta-learning using CNN on the CWRU and Paderborn University (PU) dataset. The study also focused on classifying the bearing faults using a limited amount of data. Other than these two datasets, a study by Vos et al. (Vos et al., 2022) employed AD for vibration-based fault diagnosis. Experimented on Airbus and DST gearbox datasets, the study employed LSTM-SVM and simple OCSVM techniques.

For this research, we have used seven AD algorithms, including traditional approaches like iF (Liu, Ting, & Zhou, 2008), Local Outlier Factor (LOF) (Breunig, Kriegel, Ng, & Sander, 2000), one class support vector machine(OCSVM) (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 1999), and the Deep Learning (DL)-based techniques like Autoencoder (AE) (Ahmad, Styp-Rekowski, Nedelkoski, & Kao, 2020) and Variational AE (VAE) (Zhang, Ye, Wang, & Habetler, 2019), and the hybrid approaches like ResNet (He et al., 2016) and VGGNet (Simonyan & Zisserman, 2014)-based iF, LOF and OCSVM, which will be described in detail in later sections. The motive behind taking the traditional algorithms is that, for fault or anomaly detection, it is not necessarily true that DL architectures are always superior (Wang, Vos, et al., 2023; Audibert, Michiardi, Guyard, Marti, & Zuluaga, 2022). The traditional algorithms, with the simpler architectures, can sometimes outperform the complex and deeper networks.

The organization of this article is as follows. In Section 2, the dataset description is presented. Section 3 provides an overview of the methodology implemented in this research, and Section 4 presents the experimental results and analysis of this work. Finally, the article concludes in Section 5.

2. Dataset Description

We have used three datasets for this research, two of which are the most widely used benchmark datasets—the CWRU and PU bearing datasets— and the other is the Health and Usage Monitoring System (HUMS) planet gear rim crack dataset provided by the Defence Science and Technology Group (DSTG) in Melbourne, Australia.

2.1. CWRU Dataset

The CWRU bearing dataset is one of the most widely used fundamental bearing datasets for MFD research. It contains experimental data collected from a test rig with four different types of faults: inner race fault, outer race fault, ball fault, and normal (healthy) state. These faults are artificially induced with varying severities and load conditions. The dataset provides time-domain vibration signals, making it suitable for MFD methods such as feature extraction, classification, and model training (Chaleshtori & Aghaie, 2024). The dataset is publicly available on this website [1]. For this research, we have used all four types of faults with a fault diameter of 7 mils (1 mils=0.001 inches) with all available loads from 0 to 3 HP. A total of 413 instances were used for each class. The types of faults used are shown in Table 1.

2.2. PU Dataset

The PU dataset, provided by the KAT data center at Paderborn University, is a comprehensive resource for MFD and prognosis research. The PU bearing dataset comprises vibration data from experiments on six healthy bearings and 26 damaged bearing sets, of which 12 are artificial damages, and 14 are real damages. The dataset provides time-domain vibration signals, acoustic emission signals, and temperature measurements, covering various fault severities and load conditions (Lessmeier, Kimotho, Zimmer, & Sextro, 2016; Neupane, Bouadjenek, Dazeley, & Aryal, 2024). This dataset can be downloaded from this website [2]. For this research, we have taken five types of bearing vibration data, including two artificial fault types, two real fault types, and one normal state data. A total of 4967 instances from each class were used. Other information about the dataset is described in Table 1.

2.3. HUMS Dataset

The HUMS dataset originates from an extensive experimental study executed at the Helicopter Transmission Test Facility (HTTF) at the DSTG in Melbourne. This study was executed with the specific aim of investigating fatigue cracking in thin-rim helicopter planet gears, which are critical components of helicopter transmission systems. The dataset was released as a part of the HUMS 2023 Data Challenge. Further information about the experimental set, data processing, and acquisition

[1] https://engineering.case.edu/bearingdatacenter
[2] https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-datacenter/

Table 1. Types of faults and number of instances used for the CWRU and PU datasets

CWRU (413)	**PU (4967)**
Normal	Normal
B007	KA01 (Artificial Damage [OR])
IR007	KA03 (Artificial Damage [OR])
OR007	KB23 (Real Damage [IR+OR])
	KB24 (Real Damage [IR+OR])
All Faults (B+IR+OR)	All Faults (Artificial+Real)

Table 2. Number of data files (records) provided for the HUMS dataset

Day	**No. of records**	**Remarks**
Day 17	65	Provided Later
Day 18	68	
Day 19	62	
Day 20	87	*Total 282*
Day 21	89	
Day 22	80	
Day 23	72	Provided Earlier
Day 24	89	
Day 25	85	
Day 26	26	*Total 526*
Day 27	27	
Grand Total	**808**	

technique for this dataset can be found on (Peeters, Wang, Blunt, Verstraeten, & Helsen, 2024), (Wang, Blunt, & Kappas, 2023), and (Sawalhi, Wang, & Blunt, 2024). A total of 808 four-channel planet-ring hunting-tooth average data files were provided in two sessions (526 files [*files from Day 21 to Day 27*] before the data challenge and 282 files [*from Day 17 to Day 20*] after the challenge). The whole dataset features 94 load cycles, out of which the last 60 cycles were released prior to the data challenge, and the first 34 load cycles were released later. Table 2 shows the number of records with respect to the days of testing. In this research, we used 282 data files from Day 17 to Day 20, which were taken as a training set, and the remaining 526 data files from Day 21 to Day 27 were taken as the test set. Our experiment encompassed data collected from all four sensors.

3. Methodology Implemented

The methodology implemented in this research is consistent across two benchmark datasets, CWRU and PU, with a minor difference in the pre-processing (PP) step for the HUMS dataset.

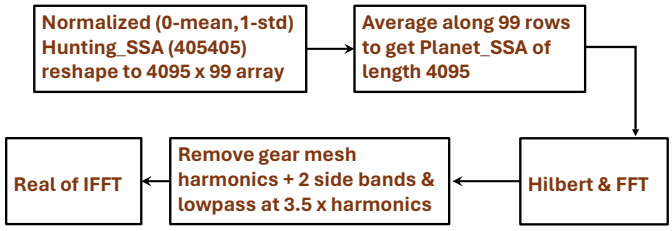


Figure 4. Preprocessing technique used for the HUMS dataset

3.1. Pre-processing

A. ***CWRU and PU datasets:*** The initial preprocessing step of standardizing the raw vibration signals was done to achieve a mean of zero and a standard deviation of one. Then, the signals of length, X (say), were segmented into N samples, each comprising 4096 data points. It is important to note that the value of N varies across datasets but remains constant for different fault types within a particular dataset.

B. ***HUMS dataset:*** The whole dataset consists of 808 files of Hunting tooth synchronous averaging (H-SSA) with 405405 data points per sample per channel, which was standardized to zero mean and unit standard deviation. This standardization of H-SSA mitigates variations in torque, speed, and temperature, enhancing sensitivity to fault-induced changes. Then, Planet Gear SSA (P-SSA) was derived by reshaping H-SSA into a matrix and averaging along specific rows corresponding to gear revolutions. Specifically, each 405405-data points sample was reshaped into a 4095×99 matrix array and was averaged along 99 rows to get the averaged sample of 4095 data points. The output 4095 data points sample was then transformed using Hilbert and then fast Fourier transform. The residual signals were generated by eliminating gear mesh harmonics and sidebands in the order domain. To detect rim cracks, an ideal low-pass filter at 3.5 times the gear mesh harmonics was applied, followed by an inverse fast Fourier transform (IFFT), and the real values of IFFT were taken as the data points for samples (Sawalhi et al., 2024), (Peeters et al., 2024). In this way, 808 planet-ring hunting-tooth average samples per channel, each of length 4095, were finally achieved. The preprocessing steps for the HUMS dataset can also be seen in figure 4.

3.2. Analyses Carried

After these prepossessing steps, two primary analyses were conducted for all three datasets:

A. ***Statistical analysis:*** For each segment generated, key statistical metrics including, Mean (M), Standard deviation (Std), Peak-to-Peak (P2P), Kurtosis (K), and Skewness (Sk) were computed and saved in a *CSV* format. Furthermore, labels were assigned to each of the samples of the

CWRU and PU datasets to indicate their condition, with '0' representing a normal state and '1' signifying a fault. Since the HUMS dataset does not contain distinctive normal and faulty signals, this labelling step was skipped for this particular dataset.

B. ***Wavelet transform analysis:*** Scalograms were generated from the pre-processed data files for each datasets, for further examination using the Continuous Wavelet Transform (CWT) (Zheng, Li, & Chen, 2002) technique, specifically employing the Morlet wavelet. Research (Neupane, Kim, & Seok, 2021), (Guo, Liu, Li, & Wang, 2020) indicate that vibration signals featuring periodic impulses correspond notably with the Morlet wavelet's properties. This alignment facilitates the utility of Morlet wavelets in identifying both anomalies and standard elements in machinery, which has made it a popular choice in this domain of study. Scalograms were labelled as '0' or '1' to indicate normal or faulty signals for the CWRU and PU datasets, and skipped for the HUMS dataset.

3.3. Anomaly Detection Approaches

Anomalies represent data instances that exhibit distinct characteristics from normal instances, and the detection of these abnormal patterns or instances is called anomaly detection (Liu et al., 2008). AD, also called outlier detection, is a widely used technique in data mining and ML to identify or detect instances or patterns that do not conform to the expected behavior within a dataset (Kumagai, Iwata, & Fujiwara, 2021). AD methods have been used in various applications, such as fraud detection (Pourhabibi, Ong, Kam, & Boo, 2020), intrusion detection (Aryal, Santosh, & Dazeley, 2021), and so on. The task of AD can be addressed through supervised, semi-supervised, or unsupervised learning strategies. However, a significant obstacle is the scarcity of high-quality training instances, particularly for anomalous behaviors, which pose challenges in various domains, including MFD. Given these challenges, it is imperative to address the task through semi-supervised approaches.

Semi-supervised AD techniques are designed to identify anomalies or outliers in data by combining labelled and unlabelled instances. The process begins by manually labeling a small subset of the data as either normal or anomalous, which serves as the training set. Using this labelled data, a model is trained to distinguish between these two categories. Subsequently, the trained model is applied to the unlabelled data, assigning scores or probabilities to each data point. Thresholds are then applied to these scores to classify instances as either normal or anomalous.

For this work, we have labelled only the normal data and trained the AD models on this subset of labelled data. We explored the efficacy of various AD algorithms like iF, LOF, OCSVM, AE, and VAE. The use of statistical features is primarily for traditional AD algorithms, like iF, LOF, and OCSVM only. In contrast, the scalogram images are fed as input to the DL architectures, like AE, and VAE. Additionally, DL architectures like ResNet50 and VGG16 are employed to extract the features from the scalograms, and traditional algorithms (iF, LOF, and OCSVM) are employed for the extracted features for detecting normal and anomalous instances. A brief overview of each of these algorithms is provided below:

- **iF:** Isolation forest (Liu et al., 2008) is an AD algorithm that operates on a tree-based approach to identify outliers in the dataset. This algorithm isolates anomalies by randomly selecting features and partitioning data points based on their values along those features. This process is repeated recursively until each data point is isolated in its own partition. Anomalies are identified as data points that require fewer partitions to isolate, as they stand out as unusual compared to normal instances.
- **LOF:** Local outlier factor (Breunig et al., 2000) is a density-based AD algorithm, that measures the local deviation of a data point in relation to its neighbors. It calculates the ratio of the local density of a point to the local densities of its neighbors, identifying outliers as data points with significantly lower densities compared to their neighbors.
- **OCSVM:** One class support vector machines (Schölkopf et al., 1999), an AD algorithm used for novelty detection, constructs a hyperplane that separates the normal data instances from the origin in a high-dimensional feature space. This method aims to maximize the margin between the hyperplane and the nearest normal data points, identifying anomalies as data points lying on the opposite side of the hyperplane from the normal class.
- **AE:** Autoencoders (Torabi, Mirtaheri, & Greco, 2023), a type of neural network architecture, can also be used for AD tasks. When trained on normal data points, AE aims to reconstruct input data with minimal error; however, anomalies generally result in higher reconstruction errors. By setting a predefined threshold, instances with reconstruction errors surpassing this threshold are flagged as anomalies or outliers.
- **VAE:** Variational AEs (Xie, Xu, Jiang, Gao, & Wang, 2024), a variation of AE, are capable of learning complex data distributions and generating new data samples similar to the training data. VAEs, trained on normal data points, aim to reconstruct input data with minimal error. However, anomalies typically result in higher reconstruction errors, as they deviate significantly from the learned data distribution. By comparing the reconstructed data with the original input, anomalies can be identified based on higher reconstruction errors.

Moreover, we have also used ResNet50 (He et al., 2016), and VGG16 (Simonyan & Zisserman, 2014) neural architectures for feature extraction from the scalogram images. These

are pre-trained architectures, which utilize a series of convolutional and pooling layers to extract hierarchical features from input images. ResNet50 introduces residual connections, which help alleviate the vanishing gradient problem during training, allowing for deeper architectures to be trained effectively. In contrast, VGG16 relies on a simpler architecture with a stack of convolutional layers followed by max-pooling layers. Despite the difference in their architecture, both of these networks can extract informative features from images. The extracted features are used as the input of three AD models: iF, LOF and OCSVM.

Thus, the methodology incorporates three diverse strategies for anomaly detection, specifically designed for those data types and analytical approaches. These approaches utilize a consistent evaluation framework, which comprises multiple runs (10), incorporates statistical and deep features, and employs various thresholding techniques for detecting anomalies. The following provides a brief overview of each approach:

A. ***Approach 1: AD with Statistical Features:*** This study evaluates the effectiveness of the key statistical features, like mean, standard deviation, kurtosis, skewness, and P2P, computed for each standardized sample, and the traditional AD algorithms in detecting anomalies. Three models, iF, LOF, and OCSVM, were implemented. A comprehensive analysis was conducted across 31 combinations of these features to explore their effectiveness in AD. The anomaly score generated by these models was compared with the custom thresholds like three sigma ($\mu - 3\sigma$), one percent, and minimum anomaly score + standard error.

B. ***DL-based End-to-End AD:*** The second strategy utilized end-to-end DL models, specifically AE and VAE, which are designed for scalogram images. This method employs reconstruction loss as a measure for AD. Anomalies are expected to have a larger reconstruction loss. The same thresholding techniques are applied to the reconstruction loss to differentiate between normal and anomalous instances. This approach explores the ability of AE and VAE to capture and reconstruct the intricate patterns present in scalogram images.

C. ***Hybrid Approach (DL + Traditional AD):*** The third methodology expands the analysis of scalogram images by employing feature extraction through the use of pre-trained DL architectures like ResNet50 and VGG16 neural networks. Similar to the first approach, the models iF, LOF, and OCSVM are implemented to the extracted features to get the anomaly scores, and the anomalies were detected utilizing the same thresholding techniques. Employing ResNet50 as a feature extractor, each image results in a feature vector of size 2048, and using VGG16, each input image results in a feature vector of size 512. These features are then fed as the input of the AD models.

3.4. Threshold Techniques

The AD algorithms generate the anomaly scores. Anomaly scores in iF are typically calculated based on the number of splits required to isolate each data point in a decision tree. Data points that require fewer splits to isolate are considered more anomalous and receive higher anomaly scores. Therefore, lower anomaly scores indicate normal behavior, while higher scores indicate anomalies. Similarly, LOF computes anomaly scores by comparing the local density of data points around each point to the density of its neighbors. Points with significantly lower density compared to their neighbors are assigned higher anomaly scores. Thus, higher LOF scores denote more anomalous behavior. Similarly, anomaly scores in OCSVM are determined based on the distance of each data point from the boundary of the region containing normal data points. Points lying farther away from this boundary are considered more anomalous and receive higher anomaly scores.

Three custom thresholds are used for this research: three sigma, one percent, and the minimum anomaly score (or reconstruction loss) plus the standard error. For $\mu - 3\sigma$, the mean of these scores (μ) is calculated, along with their standard deviation (σ). The $\mu - 3\sigma$ threshold is then determined by subtracting three times the standard deviation (3σ) from the mean ($\mu - 3\sigma$). This threshold serves as a boundary for identifying anomalies; samples with anomaly scores exceeding this threshold are considered anomalous. Additionally, for models such as AE and VAE, the reconstruction errors of normal training samples are used instead of anomaly scores. The $\mu - 3\sigma$ threshold is calculated in the same manner, but based on these reconstruction errors, providing a consistent criterion for anomaly detection across different types of models. Moreover, the one percent threshold is determined by selecting the value below which only one percent of the normal training scores or reconstruction errors fall. This threshold is established to identify anomalies among samples with exceptionally low scores, indicating significant deviations from the norm. Furthermore, the minimum value plus the standard error threshold is calculated by adding the standard error to the minimum normal training score or reconstruction error. The standard error provides a measure of the variability or uncertainty associated with the estimation of the minimum value. This threshold aims to capture anomalies beyond the minimum score while accounting for potential fluctuations.

3.5. Evaluation Framework

A. ***CWRU and PU datasets:*** The methodology follows a consistent evaluation framework across all approaches. Initially, the training data is split evenly into two halves. One half is utilized for model training, while the other half is combined with 90% of randomly selected test data to establish a diverse testing scenario. The test data includes various types of bearing health datasets collected from

the CWRU dataset, each comprising 413 instances. We created a total of five datasets, as depicted in Table 1. The 'All Faults' dataset is the combination of all fault types, namely B007, IR007, and OR007, excluding the Normal type, resulting in 1239 instances.
Additionally, we extracted five distinct health states from the PU bearing dataset. These states encompass a normal state, two artificial damages featuring OR faults, and two real damages featuring IR+OR faults, with each class containing 4967 instances. Consequently, a total of six datasets were generated, as illustrated in Table 1, in which the 'All Faults' dataset comprises all four faulty states datasets (except the normal).

B. ***HUMS Dataset:*** After the PP of the HUMS dataset, as mentioned in section 3.1, the resulting 808 data samples from each of the four sensors, were divided into train and test sets. As mentioned in an earlier section, 282 data files from the first 34 load cycles, from Day 17 to Day 20, were taken as a training set in this research, and the remaining 526 data files, from Day 21 to Day 27, were taken as the test set.

4. Experimental Results

As we have mentioned earlier, we implemented the iF, LOF, and OCSVM models which were fed with the combination of the key statistical features computed for each sample. We also employed end-to-end DL-based AD algorithms, including AE and VAE, to detect anomalies using scalogram images. Additionally, we applied ResNet50 and VGG16 architectures to extract features from the scalograms and implemented iF, LOF, and OCSVM techniques for detecting anomalies. From the experiments conducted, we obtained the following outcomes.

4.1. Results for the CWRU and PU Dataset

Tables 3 and 4 present the performance of various anomaly detection algorithms achieved for the CWRU and PU datasets, respectively. These tables represent that the feature combinations of kurtosis, skewness, and P2P excel other combinations, and the threshold $\mu-3\sigma$ performs better than other techniques. Here, the term "best average F1 score" refers to the highest F1 score calculated by averaging the F1 scores obtained from 10 separate runs. The term "Overall" denotes the best score achieved across all datasets, reflecting the highest performance observed collectively across all evaluated datasets. Abbreviations K, P2P, Sk, M and Std represent Kurtosis, Peak-to-Peak, Skewness, Mean and Standard deviation, respectively. Moreover, the average F1 score over 10 runs for each of the datasets for each method is shown as a bar graph in Figure 5 and 6. The first three bar clusters, representing models iF, LOF, and OCSVM, denote the use of the respective AD models for the feature combinations kurtosis, P2P, and skewness. The subsequent bar clusters, from ResNet-iF to VAE, use the scalogram images as input. The threshold for all of these models is $\mu - 3\sigma$. Figure 5 illustrates notable performance trends of the ResNet-iF and ResNet-OCSVM models across all dataset types for the CWRU dataset, whereas figure 6 illustrates notable performance trends of ResNet-OCSVM models across all dataset types for PU dataset.

Table 3. Experimental results for the CWRU Dataset.

Dataset	CWRU
Model	iF
Best Average F1 Score	0.99826221 (OR007)
Overall	K, P2P, Sk; $\mu - 3\sigma$
Model	OCSVM
Best Average F1 Score	0.0.997340705 (OR007)
Overall	K, Sk, P2P; Min+stdError and $\mu - 3\sigma$
x	LOF
Model	
Best Average F1 Score	0.788509613 (B007)
Overall	$\mu - 3\sigma$
Model	ResNet-iF
Best Average F1 Score	0.995008449 (OR007)
Threshold	$\mu - 3\sigma$
Model	ResNet-LOF
Best Average F1 Score	0.8 (B007)
Threshold	$\mu - 3\sigma$
Model	ResNet-OCSVM
Best Average F1 Score	0.993120206 (All Faults)
Threshold	$\mu - 3\sigma$
Model	VGG-iF
Best Average F1 Score	0.908164235(IR007)
Threshold	One Percent
Model	VGG-LOF
Best Average F1 Score	0.8(All Faults)
Threshold	$\mu - 3\sigma$
Model	VGG-OCSVM
Best Average F1 Score	0.8(All Faults)
Threshold	$\mu - 3\sigma$
Model	AE
Best Average F1 Score	0.753205267(All Faults)
Threshold	$\mu + 3\sigma$
Model	VAE
Best Average F1 Score	0.872920403(All Faults)
Threshold	$\mu + 3\sigma$

4.2. Results for HUMS Dataset

The HUMS dataset is a new dataset in the study of machinery faults, and researchers are employing various algorithms to detect the faults and find anomalous patterns in them. There aren't any concrete results yet. In the results of the data

Table 4. Experimental results for the PU Dataset.

Dataset	PU
Model	iF
Best Average F1 Score	0.98707402 (Artificial Damages)
Overall	K, P2P, Sk; $\mu - 3\sigma$
Model	LOF
Best Average F1 Score	0.935198014 (All Faults)
Overall	Sk; $\mu - 3\sigma$
Model	OCSVM
Best Average F1 Score	0.985556437(Artificial Damages)
Overall	K, P2P, Std; $\mu - 3\sigma$ and Min+stdError
Model	ResNet-iF
Best Average F1 Score	0.930936511(Real Damages)
Threshold	$\mu - 3\sigma$
Model	ResNet-OCSVM
Best Average F1 Score	0.999316099 (All Faults)
Threshold	$\mu - 3\sigma$ and Min+stdError
Model	VGG-iF
Best Average F1 Score	0.981120622(Real Damages)
Threshold	$\mu - 3\sigma$
Model	VGG-OCSVM
Best Average F1 Score	0.941165324 (All Faults)
Threshold	$\mu - 3\sigma$

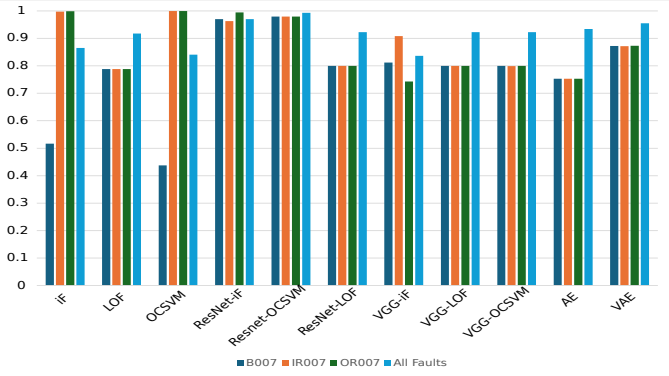


Figure 5. Comparison of Models' Performances on the CWRU Dataset.

challenge, the winning team (Peeters et al., 2024) claimed the record number #175 (Day 23/ 20211214, 104944) to be the earliest convincing fault detection. However, the data challenge committee pointed out that records #264 (Day 24/ 20211216, 112716) and #272 (Day 24/ 20211216, 120021) as contenders. As further research continues, different results are claimed, proposing different records as the earliest detection. In the latest notice released by the committee[3], records #15 (Day 21/ 20211208, 113917), #50 (Day 21/ 20211208, 135820), #125 (Day 22/ 20211209, 124241), #143 (Day 22/ 20211209, 135146) and #150 (Day 22/ 20211209, 141330) have found to contain the anomalies as well.

[3]https://www.dst.defence.gov.au/our-technologies/helicopter-main-rotor-gearbox-planet-gear-fatigue-crack-propagation-test

Figure 6. Comparison of Models' Performances on PU Dataset

With our various AD detection algorithms, various records (or file numbers) were detected as the earliest detection. However, seeing the most convincing features (kurtosis, P2P, and skewness) and effective algorithms for the CWRU and PU dataset, the results obtained from ResNet-iF and ResNet-OCSVM are considered for this HUMS dataset as well. The iF, LOF and OCSVM algorithms, trained on the combined features of kurtosis, skewness and P2P and threshold $\mu - 3\sigma$, predicted #15 (Day 21/ 20211208, 113917), #50 (Day 21/ 20211208, 135820) and #150 (Day 22/ 20211209, 141330) as the first three consecutive faults. Taking the ResNet-iF and ResNet-OCSVM models and $\mu - 3\sigma$ as a threshold, the earliest anomaly prediction was found to be the file #11 (Day 21/20211208, 112723).

5. Discussion and Conclusion

Identifying faults in machinery poses significant challenges, particularly in accurately classifying fault types. Conventional supervised machine learning methods have limitations due to the need for abundant labelled data, expert supervision in labelling, and their inability to generalize to unseen faults. To tackle these challenges, this article explores the potential of semi-supervised learning-based anomaly detection techniques in the field of machinery fault diagnosis. This study specifically focuses on identifying abnormal patterns in machinery vibration signals, which are crucial for preventing breakdowns and ensuring safety and productivity. Our experimental results highlight the effectiveness of certain feature combinations, such as kurtosis, skewness, and peak-to-peak, in conjunction with a threshold of three sigma. Furthermore, we found that models like ResNet-OCSVM and ResNet-iF, as well as deep learning-based methods like VAE, demonstrate promising performance. However, it's worth noting that DL-based techniques often come with higher computational resource requirements and longer training times, as depicted

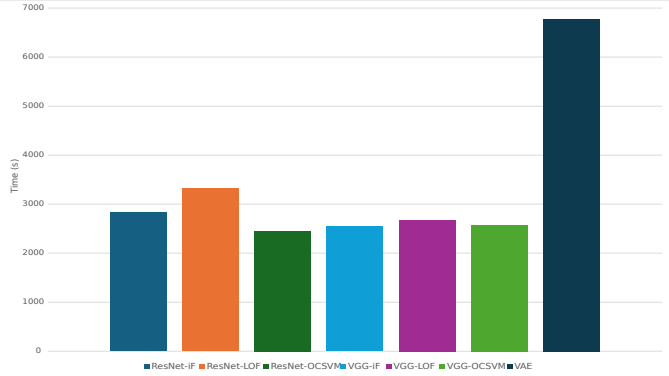


Figure 7. Comparison of Model Performances Based on Runtime: The figure illustrates the time taken by various models to complete 10 runs of anomaly detection using all 4 test sets from the CWRU dataset. Notably, all models operate on the same input, namely, scalograms.

in figure 7. Interestingly, simpler traditional methods, sometimes, outperform or perform equally well compared to complex DL methods. Given their simplicity and lower computational demands, prioritizing these simpler approaches may be more practical in many scenarios.

Our research examines seven AD methods across various feature representations using benchmark datasets, including the CWRU bearing, PU bearing, and HUMS planet gear rim crack dataset. Our findings provide valuable insights with significant practical implications, suggesting that simpler approaches may be, sometimes, effective in real-world applications due to their ease of implementation and reduced computational burden. DL methods, indeed, have shown promising results in MFD, but their practicality may be limited by resource constraints. Therefore, incorporating semi-supervised learning-based AD techniques alongside simpler traditional methods can enhance fault detection systems in industrial settings. We, therefore, would like to recommend that future researchers proceed with simpler methods initially, then transition to DL-based methodologies if necessary for MFD.

References

Ahmad, S., Styp-Rekowski, K., Nedelkoski, S., & Kao, O. (2020). Autoencoder-based condition monitoring and anomaly detection method for rotating machines. In *2020 ieee international conference on big data (big data)* (pp. 4093–4102).

Aryal, S., Santosh, K., & Dazeley, R. (2021). usfad: a robust anomaly detector based on unsupervised stochastic forest. *International Journal of Machine Learning and Cybernetics*, *12*, 1137–1150.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2022). Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, *132*, 108945.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 acm sigmod international conference on management of data* (pp. 93–104).

Chaleshtori, A. E., & Aghaie, A. (2024). A novel bearing fault diagnosis approach using the gaussian mixture model and the weighted principal component analysis. *Reliability Engineering & System Safety*, *242*, 109720.

Das, O., Das, D. B., & Birant, D. (2023). Machine learning for fault analysis in rotating machinery: A comprehensive review. *Heliyon*.

Guo, J., Liu, X., Li, S., & Wang, Z. (2020). Bearing intelligent fault diagnosis based on wavelet transform and convolutional neural network. *Shock and Vibration*, *2020*, 1–14.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Jiao, J., Zhao, M., Lin, J., & Liang, K. (2019). Hierarchical discriminating sparse coding for weak fault feature extraction of rolling bearings. *Reliability Engineering & System Safety*, *184*, 41–54.

Kumagai, A., Iwata, T., & Fujiwara, Y. (2021). Semi-supervised anomaly detection on attributed graphs. In *2021 international joint conference on neural networks (ijcnn)* (pp. 1–8).

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *Phm society european conference* (Vol. 3).

Li, C., Zhang, S., Qin, Y., & Estupinan, E. (2020). A systematic review of deep transfer learning for machinery fault diagnosis. *Neurocomputing*, *407*, 121–135.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (p. 413-422). doi: 10.1109/ICDM.2008.17

Neupane, D., Bouadjenek, M. R., Dazeley, R., & Aryal, S. (2024). Data-driven machinery fault detection: A comprehensive review. *arXiv preprint arXiv:2405.18843*.

Neupane, D., Kim, Y., & Seok, J. (2021). Bearing fault detection using scalogram and switchable normalization-based cnn (sn-cnn). *IEEE Access*, *9*, 88151-88166. doi: 10.1109/ACCESS.2021.3089698

Neupane, D., & Seok, J. (2020). Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review. *IEEE Access*, *8*, 93155–93178. doi: 10.1109/ACCESS.2020.2990528

Peeters, C., Wang, W., Blunt, D., Verstraeten, T., & Helsen, J. (2024). Fatigue crack detection in planetary gears:

Insights from the hums2023 data challenge. *Mechanical Systems and Signal Processing*, *212*, 111292.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303.

Sawalhi, N., Wang, W., & Blunt, D. (2024). Helicopter planet gear rim crack diagnosis and trending using cepstrum editing enhanced with deconvolution. *Sensors*, *24*(8), 2593.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. *Advances in neural information processing systems*, *12*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Torabi, H., Mirtaheri, S. L., & Greco, S. (2023). Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity*, *6*(1), 1.

Vos, K., Peng, Z., Jenkins, C., Shahriar, M. R., Borghesani, P., & Wang, W. (2022). Vibration-based anomaly detection using lstm/svm approaches. *Mechanical Systems and Signal Processing*, *169*, 108752.

Wang, W., Blunt, D., & Kappas, J. (2023). *Helicopter main gearbox planet gear crack propagation test dataset.*

Wang, W., Vos, K., Taylor, J., Jenkins, C., Bala, B., Whitehead, L., & Peng, Z. (2023). Is deep learning superior to traditional techniques in machine health monitoring applications. *The Aeronautical Journal*, *127*(1318), 2105–2117.

Xie, T., Xu, Q., Jiang, C., Gao, Z., & Wang, X. (2024). A robust anomaly detection model for pumps based on the spectral residual with self-attention variational autoencoder. *IEEE Transactions on Industrial Informatics*.

Zhang, S., Ye, F., Wang, B., & Habetler, T. G. (2019). Semi-supervised learning of bearing anomaly detection via deep variational autoencoders. *arXiv preprint arXiv:1912.01096*.

Zhang, S., Ye, F., Wang, B., & Habetler, T. G. (2020). Few-shot bearing anomaly detection via model-agnostic meta-learning. In *2020 23rd international conference on electrical machines and systems (icems)* (p. 1341-1346). doi: 10.23919/ICEMS50442.2020.9291099

Zhang, S., Ye, F., Wang, B., & Habetler, T. G. (2021). Semi-supervised bearing fault diagnosis and classification using variational autoencoder-based deep generative models. *IEEE Sensors Journal*, *21*(5), 6476-6486. doi: 10.1109/JSEN.2020.3040696

Zheng, H., Li, Z., & Chen, X. (2002). Gear fault diagnosis based on continuous wavelet transform. *Mechanical systems and signal processing*, *16*(2-3), 447–457.

Zhong, X., Zhang, L., & Ban, H. (2023, May). Deep reinforcement learning for class imbalance fault diagnosis of equipment in nuclear power plants. *Annals of Nuclear Energy*, *184*, 109685. doi: 10.1016/j.anucene.2023.109685

Zong, X., Yang, R., Wang, H., Du, M., You, P., Wang, S., & Su, H. (2022). Semi-supervised transfer learning method for bearing fault diagnosis with imbalanced data. *Machines*, *10*(7). doi: 10.3390/machines10070515

A Computer Vision Deep Learning Tool for Automatic Recognition of Bearing Failure Modes

Stephan Baggeröhr[1], Sebastián Echeverri Restrepo [1,2], Mourad Chennaoui [1], Christine Matta [1] and Cees Taal[1],

[1] *SKF, Research and Technology Development Center, Houten, the Netherlands*
stephan.baggerohr@skf.com
cees.taal@skf.com
sebastian.echeverri.restrepo@skf.com
mourad.chennaoui@skf.com
christine.matta@skf.com

[2] *Department of Physics, King's College London, London, United Kingdom*

ABSTRACT

We introduce an object detection model specifically designed to identify failure modes in images of bearing components, including the inner ring, outer ring, and rolling elements. The method effectively detects and pinpoints failure features, subsequently determining the associated failure mode within the image. With images sourced from real-world bearing applications, our model can recognize various ISO-failure modes such as surface-initiated fatigue, abrasive wear, adhesive wear, moisture corrosion, fretting corrosion, current leakage erosion, and indents from particles. The proposed model could be used in an assistive tool where failure modes give insights on how to prolong average future bearing life in an asset and therefore reduce related costs and environmental impacts.

1. INTRODUCTION

Bearings are extensively utilized in a wide range of rotating equipment and are essential for ensuring their proper function. Bearing failures can lead to unplanned downtime with unforeseen costs, or even result in hazardous situations. Sensor-based condition monitoring has been an important tool for the prediction of these undesired events and are a key ingredient for a predictive maintenance strategy (Randall & Antoni, 2011). In this paper, the focus is on a subsequent stage after sensor-based fault detection, that is, a visual inspection of the replaced disassembled bearing to further prolong the average future bearing life in an asset (SKF, 2017).

A visual inspection of the bearing provides additional information on how to prevent problems from reoccurring. This includes altering the bearing design, lubrication or operation and maintenance procedures. Another important application of bearing inspections is quantifying its damage severity, such as spall size. This information can be fed back to sensor-based condition monitoring systems enabling supervised machine learning for bearing diagnostics and prognostics. Inspections are also being used to determine whether a bearing qualifies for remanufacturing (Chiarot, Cooper Ordoñez, & Lahura, 2022). Remanufacturing is a process which enables re-using the bearing by means of polishing or grinding its components, potentially doubling its life. To summarize, visual bearing inspections can significantly prolong average future bearing life in an asset and therefore reduce related costs and environmental impacts, e.g., due to the manufacturing process of the bearing.

Unfortunately, visual postmortem analysis of bearings require an application engineer with many years of experience, which is something not always readily available. This limits its scalability to be applied to a large population of bearings used in an asset. In this work we propose an image-based deep learning algorithm, which can assist the technician replacing the bearing. For example, a picture can be taken of the bearing components with a smart-phone, where the software automatically provides insights on proposed maintenance actions, altering bearing designs, its remanufacturability and provide an automated connection in supervising condition monitoring algorithms.

Bearing failures can occur due to a wide variety of reasons (Liu & Zhang, 2020), where each failure category can lead to a unique footprint observable during visual inspection (SKF, 2017). The different categories of bearing failures have been standardized and well described in (ISO-15243-2017, 2017), also referred to as bearing failure modes, where, in total,

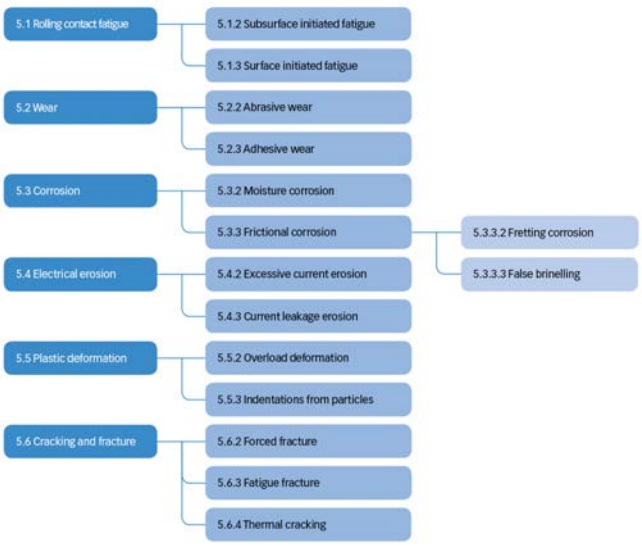


Figure 1. *ISO 15243-2017 failure mode classifications. Image taken from (SKF, 2017).*

seven main categories of failure modes are proposed. An overview of the different failure modes is shown in Figure 1. In Figure 2 an overview is shown on the most common failure modes based on collected statistics from bearing inspections (SKF, 2017).

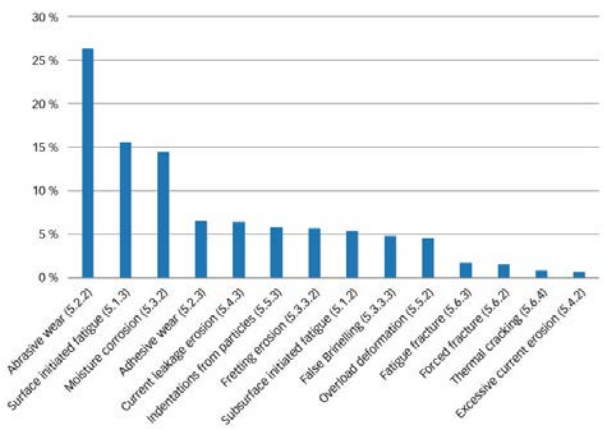


Figure 2. *An example of SKF's field failure statistics, detailing the frequency of various failure modes. Image taken from (SKF, 2022).*

Applying deep learning algorithms to automate visual inspections in PHM applications is not new. A significant amount of work has been done in the field of structure health monitoring. Examples include crack detection in concrete structures caused by, e.g., changing loading and corrosion (Azimi, Eslamlou, & Pekcan, 2020). More examples can be found from the steel industry, that is, detection and classification of steel surface defects (Fu et al., 2019; Wang, Xia, Ye, & Yang, 2021). However, to the authors knowledge there is no specific method to classify bearing failure modes.

In this work a framework of selecting a deep-learning based object detection model is introduced. The object detection model is specifically tasked to identify failure modes in images of bearing components, including the inner ring, outer ring, and rolling elements. This model effectively detects and pinpoints failure features, subsequently determining the associated failure mode within the image. As a first step, the selected model is trained for the top 7 most common failure modes, namely: abrasive wear, surface-initiated fatigue, moisture corrosion, adhesive wear, current leakage erosion, fretting corrosion, and indentations from particles (Figure 2).

2. Dataset

The foundation of bearing failure mode object detection model lies in the curated dataset. The dataset encompasses a broad spectrum of bearing images, taken from industrial centres across the globe and showcases various bearing types along with the one or more of the top seven primary failure modes identified for diagnosis. This breadth in dataset variety was crucial for the development of a model capable of accurately identifying and classifying a range of real-world bearing failures captured in their operational environment.

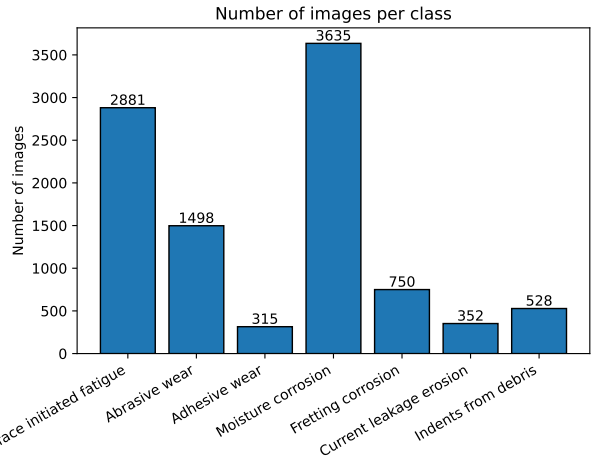


Figure 3. *Final number of images per failure class after selecting process and annotation done by expert.*

The precision in our dataset was ensured by an expert led data labelling team based on SKF employee's experience. Specialists in bearing maintenance meticulously labeled and annotated each image, drawing accurate bounding boxes around the designated failure modes. During the annotation process images were selected based on their representation of the failure mode, making sure the failure mode characteristics and features are within clear view according to the ex-

pert. Furthermore, an assessment was made on the quality of the image itself, filtering any blurry images. Images objects other than bearing components (maintenance tools, tables, etc) were also removed from the training set. In the end the dataset comprised of 11k images across the 7 chosen failure modes as shown in Figure 3. Images were normalized, padded and resized to 640x640 pixels. Additionally, augmentation techniques were applied to the images before ingesting into the model.

Here the bar graph illustrates a significant class imbalance within our object detection dataset, where certain classes are overly represented with a high number of images, while others have markedly fewer instances. This imbalance poses a challenge for effective model training, as it can lead to biases towards the more prevalent classes, potentially compromising the model's ability to accurately detect and classify less represented objects. Addressing this issue is crucial for enhancing the model's overall performance and ensuring a balanced sensitivity across all classes. To overcome this, as a first step, the shift-scale-rotate augmentation was applied with a rotation limit set to +/- 15 degrees. This method involves stochastic affine transformations that adjust the original images through shifting, scaling, and rotating. Such transformations significantly increase the dataset's variety without the need to collect new samples.

3. Proposed Methodology

The methodology employed in developing an object detection model aimed at detecting failure modes in images of bearings was twofold: firstly, leveraging out-of-the-box (pretrained) models, and secondly, fine-tuning these models on the earlier described dataset split into an 80%-20% training and test set, respectively.

3.1. Model Selection

To determine the optimal pre-trained model for our application, we conducted a comparative analysis of several state-of-the-art models. Each model was evaluated using its default parameters, with the only modifications being the image size and batch size. Specifically, all models were trained with images resized to 640x640 pixels and a batch size of 4. The models included in the study were as follows with their respective backbone (Zou, Chen, Shi, Guo, & Ye, 2023):

- EfficientDet (D0)
- EfficientDet (D4)
- Retinanet (Resnet - 101 - 2x)
- Retinanet (Resnet - 101 - 1x)
- Retinanet (Swin)
- Yolo-x (Yolo - Tiny)

The models were compared using the COCO metric. The COCO metric, used for evaluating object detection models, includes several key components: Average Precision (AP) and Average Recall (AR) across multiple IoU thresholds (0.50 to 0.95). The metric also evaluates performance across different object sizes (small, medium, large), providing a comprehensive and standardized assessment of a model's detection capabilities. This robust evaluation ensures accurate localization and detection across varied conditions (Lin et al., 2014).

3.2. Training the Model

The dataset, characterized by class imbalance among different failure modes, necessitated a tailored approach to model training. To mitigate the effects of class imbalance, focal loss was integrated into the model's loss function (Lin, Goyal, Girshick, He, & Dollár, 2017). This modification aimed to amplify the loss associated with misclassified examples, particularly those from underrepresented classes, thereby enhancing the model's sensitivity to such cases. The models were trained with a learning rate of 1e-4 for 20 epochs.

One of the paramount challenges encountered during training was the potential for overfitting. To counteract this, techniques such as early stopping, layer normalization, and weight decay were employed. Additionally, model performance was evaluated using the test set to ensure generalizability beyond the training data. Early stopping as a regularization technique was also used to prevent overfitting, by halting the training process before the model's performance on the test set starts to degrade. By terminating the training at this optimal point, early stopping ensures that the model retains its ability to generalize well to new, unseen data, thereby mitigating overfitting and improving the model's overall predictive performance.

4. Results

Figure 5 shows the results of the comparative study of different model architectures. The study revealed that EfficientDet and RetinaNet emerged as top candidates in terms of accuracy in contrast to the Yolo methods. The RetinaNet model, with its ResNet backbone, was ultimately selected based on its performance.

The implementation of early stopping mechanisms helped mitigate this risk by halting training once the test loss plateaued, as shown in Figure 6. This strategy proved invaluable in preserving the model's generalizability.

Example model predictions for the different failure modes are shown in Figure 4. Looking at the confusion matrix in Figure 7, the Retinanet model detection threshold was set in a way that left around 32% of the images without any predictions resulting in low recall. Among the images with predictions, there was a notable emphasis on precision, as evidenced by a significant number of predictions aligning along the matrix's

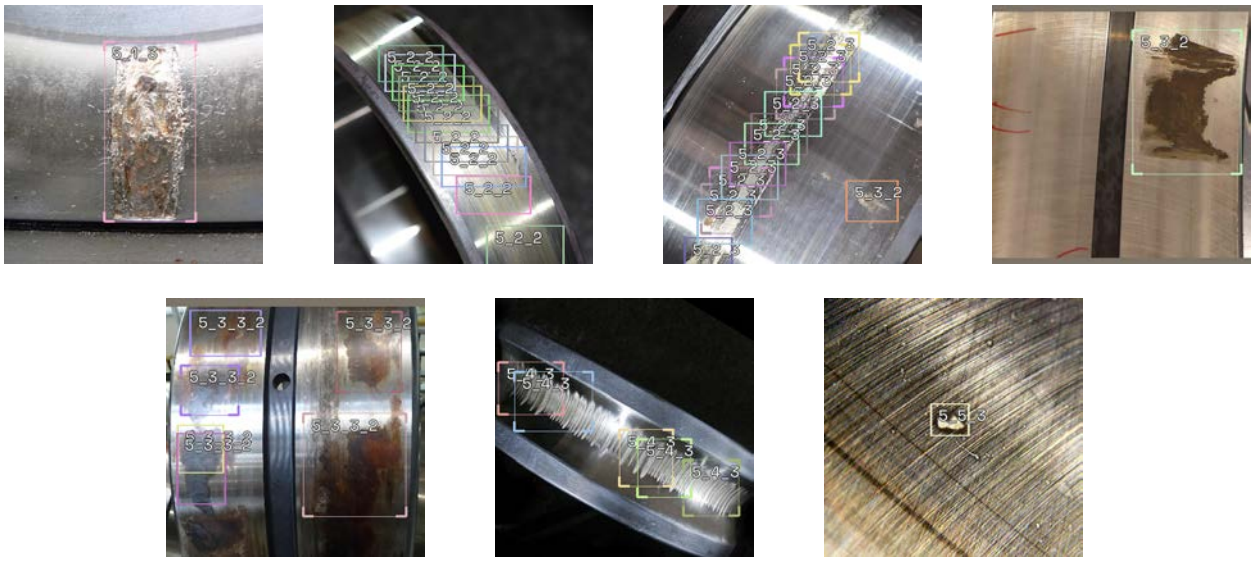


Figure 4. Example predictions for several failure modes. From left to right, top to bottom: Surface initiated fatigue, abrasive wear, adhesive wear, moisture corrosion, fretting corrosion, current leakage erosion and indentations from debris.

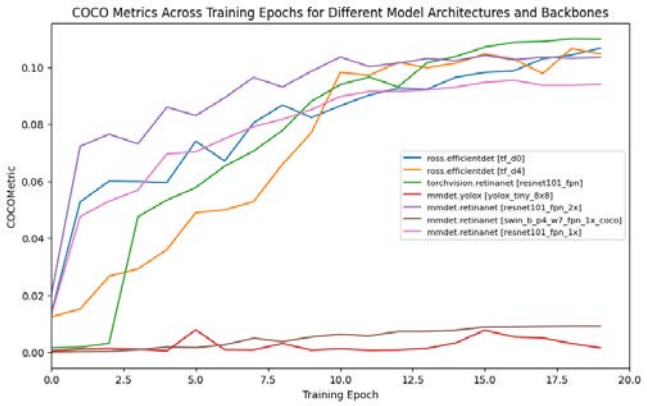


Figure 5. *Comparative analysis of investigated object detection models.*

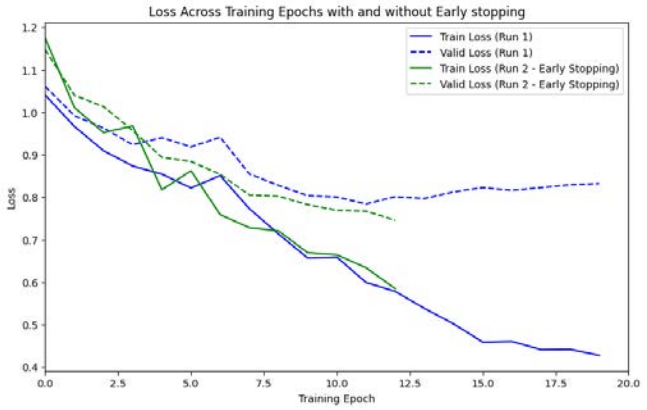


Figure 6. *Graph depicting model accuracy along epochs with and without early stopping indicated to prevent overfitting.*

diagonal.

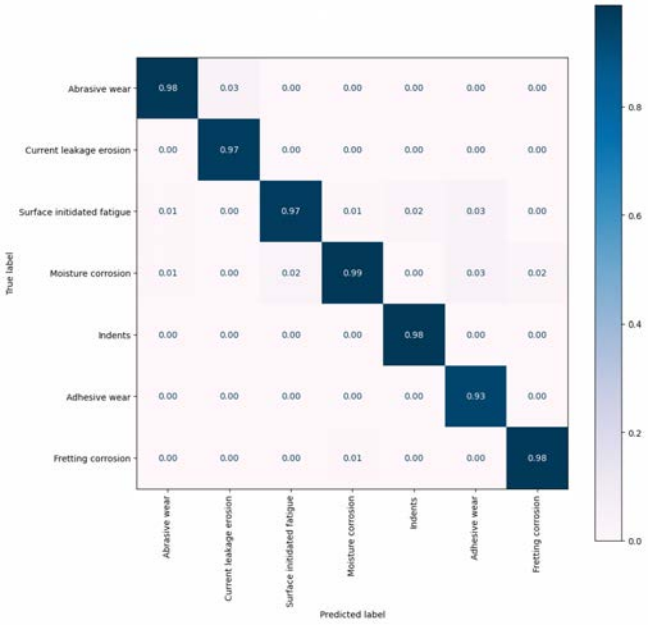


Figure 7. *Confusion matrix for the top performing model (RetinaNet - ResNet 101, fpn) model applied to the test set. Displayed results pertain exclusively to images with predictions.*

In evaluating the performance of the object detection models, we have observed a notable discrepancy between the model's precision and recall, as measured by the COCO metric system. Specifically, our model demonstrates high precision (as shown in Figure 7), indicating a strong ability to correctly identify and label objects when it decides to do so. However,

this is adjacent to a significantly lower recall, suggesting that the model is more conservative in its detection, often missing objects that should have been detected. This characteristic leads to a lower overall COCO metric score, which incorporates both precision and recall into its evaluation. Despite this, the high precision of our model still presents substantial utility in specific applications where the cost of false positives is high, and accuracy in the detection of identified objects is paramount. In such scenarios, our model's ability to minimize incorrect detection — ensuring high confidence in the positive detection it makes — can be more valuable than detecting every possible object, underscoring the importance of considering application-specific requirements when evaluating model performance. Therefore, while the overall COCO metric may be lower, the high precision of our model affirms its applicability and effectiveness in contexts where precision is critically valued over recall.

5. Conclusion

The model, selected through bench-marking various neural network architectures, was trained to detect seven primary bearing failure modes, addressing challenges such as class imbalance and image rotation inconsistencies. Key to the success was the meticulous collection and preparation of images. A dataset comprising 11k images of bearings with annotated failure modes was curated to train the model. Through thorough data gathering, precise annotation, and strategic data augmentation, we created a robust dataset that improved the accuracy of the model and real-world applicability. RetinaNet, with its ResNet 101 - fpn backbone, was chosen for its performance. This work shows the feasibility of such a model to be used in an assistive tool where failure modes give insights on how to prolong average future bearing life in an asset and therefore reduce related costs and environmental impacts.

References

Azimi, M., Eslamlou, A. D., & Pekcan, G. (2020). Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review. *Sensors*, *20*(10).

Chiarot, C., Cooper Ordoñez, R. E., & Lahura, C. (2022). Evaluation of the applicability of the circular economy and the product-service system model in a bearing supplier company. *Sustainability*, *14*(19).

Fu, G., Sun, P., Zhu, W., Yang, J., Cao, Y., Yang, M. Y., & Cao, Y. (2019). A deep-learning-based approach for fast and robust steel surface defects classification. *Optics and Lasers in Engineering*, *121*, 397-405.

ISO-15243-2017. (2017). Rolling bearings—damage and failures—terms, characteristics and causes. *BSI Standards Publication*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Liu, Z., & Zhang, L. (2020). A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement*, *149*, 107002.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—a tutorial. *Mechanical Systems and Signal Processing*, *25*(2), 485 - 520.

SKF. (2017). Bearing damage and failure analysis. *SKF Group*.

SKF. (2022). *Bearing damage analysis: iso 15243 is here to help you.* Retrieved 2023-03-01, from `https://evolution.skf.com/bearing-damage-analysis-iso-15243-is-here-to-help-you/`

Wang, S., Xia, X., Ye, L., & Yang, B. (2021). Automatic detection and classification of steel surface defect using deep convolutional neural networks. *Metals*, *11*(3).

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, *111*(3), 257–276.

A data-driven risk assessment approach for electronic boards used in oil well drilling operations

Delia-Elena Dumitru[1], Jinlong Kang[2], Alejandro Olid-Gonzalez[3] and Ahmed Mosallam[2]

[1] *SLB, Bucharest, 060201, Romania*

DDumitru2@slb.com

[2] *SLB, Clamart, 92140, France*

JKang5@slb.com

AMosallam@slb.com

[3] *SLB, Madrid, 28020, Spain*

AOlid@slb.com

ABSTRACT

To assist subject matter experts in investigating electronic failures of drilling tools, an innovative risk assessment approach for oil well drilling operations is developed that relies on synthetic time-series data to emulate environmental factors encountered downhole, explicitly focusing on temperature, shock, and vibration. The approach involves utilizing load cycle counting to extract meaningful features from each environmental channel measured by the drilling tool. The results from experiments with features related to dwell periods (dwell time and dwell damage) and load cycles (cycle means and cycle ranges) show a significant correlation between load cycle features and the risk label. Subsequently, a tree-based machine learning model is trained to label drilling operations based on synthetic data. Several models have been trained initially with comparable results. However, the advantage of using a tree-based model, specifically extra trees, is explainability and the stochastic aspect, which translates into model robustness when applied to real data. Preliminary results from a case study indicate that this new approach is highly effective in categorizing environmental risks associated with drilling operations. This risk assessment method can significantly enhance the decision-making process in investigating electronic board failures by offering reliable decision support.

1. INTRODUCTION

The drilling process has undergone a remarkable transformation in the oil and gas industry, evolving into a complex and sophisticated endeavor. This increased complexity stems directly from the necessity of accessing and extracting valuable resources hidden deep within the earth's crust. To accomplish this daunting task, the industry relies heavily on drilling tools, which are the technological cornerstones of these operations.

Drilling tools represent exceedingly intricate systems enriched with electronics, comprising a multitude of electronic boards, each meticulously designed to fulfill specialized functions of paramount significance to the success of drilling operations. These electronic boards function as the central hubs of technological operations, assuming responsibilities encompassing data acquisition, signal processing, management of control systems, and the facilitation of seamless communication (Kang et al., 2022). Thus, the reliability and performance of these electronic boards are inexorably linked to the overall effectiveness of drilling endeavors. However, the harsh operating conditions encountered downhole, including elevated temperatures, dynamic vibrations, and substantial shocks, render these boards susceptible to complex failure modes, potentially resulting in drilling operation failures. Failed drilling operations can lead to significant financial losses and environmental concerns. Therefore, health assessment and prognostics of electronic boards in drilling tools is essential to ensure that proactive maintenance is carried out in advance to

prevent drilling operations from failing.

The current health assessment and prognostics models for electronics are predominantly data-driven. For instance, physics-of-failure-based prognostics combine sensor data with models that evaluate a component's deviation from normal operation (Pecht & Gu, 2009). Another example is the use of accelerometers to monitor the response of printed circuit boards to vibrations and predict their remaining life (Gu, Barker, & Pecht, 2009). Similarly, (Vichare & Pecht, 2009) propose a technique that extracts load parameters from time-series data to estimate remaining life and assess damage. This method focuses on identifying valuable features for prognostics without requiring the storage of large volumes of data. Additionally, (Prisacaru, Gromala, Han, & Zhang, 2022) detect faults in electronic packages through the Mahalanobis distance and clarify them using a clustering technique. They also employ Echo state networks to perform degradation assessment and remaining useful life prediction. Additional literature on data-driven approaches for electronics health assessment and prognostics can be found in the following review articles: (Bhat, Muench, & Roellig, 2023), (Bhargava et al., 2020), and (Michael G. Pecht, Myeongsu Kang, 2018).

In the context of electronic boards used for oil well drilling operations, (Kale, Carter-Journet, Falgout, Heuermann-Kuehn, & Zurcher, 2014) propose a probabilistic approach that uses operational data, drilling dynamics, and historical maintenance information to predict reliability and life of electronics. (Bhatnagar, Cassou, Masry, & Mosallam, 2021) develop a data-driven fault detection approach tailored to electronic boards in intelligent remote dual-valve systems . Similarly, (V. Gupta et al., 2023) present an automatic fault detection method based on support vector machines for resistivity subsystems in Logging-While-Drilling (LWD) tools. (Sobczak-Oramus, Mosallam, Basci, & Kang, 2022) introduce a data-driven fault detection approach for transmitter subsystems in LWD tools . Finally, (Mosallam, Kang, Youssef, Laval, & Fulton, 2023) propose a data-driven fault diagnostics approach for three power supply boards in LWD tools.

Obtaining comprehensive data and corresponding labels throughout the equipment lifecycle is essential for building data-driven models for health assessment and prognostics of electronics. Subject matter experts usually derive data labels through failure investigations, but this process can be costly and time-consuming for complex equipment. Specifically, investigating electronic board failures in drilling tools requires manually examining extensive operational environment data measured by the tools. This process is labor intensive and prone to human error, making it challenging. Considering this challenge, this paper proposes an innovative risk assessment approach for oil well drilling operations to assist subject matter experts in investigating electronic failures. One of the primary advantages of this approach is its ability to harness the power of supervised learning for efficient and objective risk assessment, compared to manual inspections of operational environment data.

Literature has shown that various factors, such as temperature, humidity, vibration, dust, electrical stress, etc., affect the performance and life of electronic components (Michael G. Pecht, Myeongsu Kang, 2018). Among these factors, failures attributed to environmental conditions like temperature, humidity, and vibration constitute a significant 84% of electronic failures (Bhargava et al., 2020). Given the paramount importance of environmental factors in electronic failures, this paper seeks to develop a method to aid the subject matter experts investigate the specific environmental factors contributing to electronic failures.

However, only temperature and vibration are considered in the proposed method. We did not account for potential factors such as dust, humidity, chemicals, and radiation. This omission is because drilling tools do not typically measure these parameters for electronic boards. The physical arrangement of electronic boards within these tools inherently protects against exposure to dust, humidity, radiation, and chemicals that may be present in the wellbore. These tools are typically enclosed within robust steel tubing, shielding internal electronics from direct contact with these environmental factors. Moreover, before tool deployment, field engineers frequently introduce nitrogen into these tools, reducing the likelihood of exposure to potentially harmful substances. As a result of these protective measures and practices, the risk of electronic board damage due to dust, humidity, radiation, and chemical exposure is significantly mitigated.

The rest of this paper is structured into four sections. The first section offers a detailed presentation of the proposed method. Following that, a case study is presented. Finally, the last section summarizes the findings and suggests potential avenues for future research.

2. Proposed Method

The proposed method consists of three steps: data generation, preprocessing and feature extraction, and modelling, as illustrated in Figure 1.

2.1. Data generation

To leverage the power of supervised learning, labeled environmental data are needed. We generate synthetic time series programmatically to remove the need for expert-labeled data. Drilling tools regularly record measurements concerning the environment, specifically, temperature, shock peak values, and vibration root mean square values; therefore, in our experiment, we generate synthetic time series data that emulate drilling conditions for each of the three channels. The simulated data incorporate various sources of random-

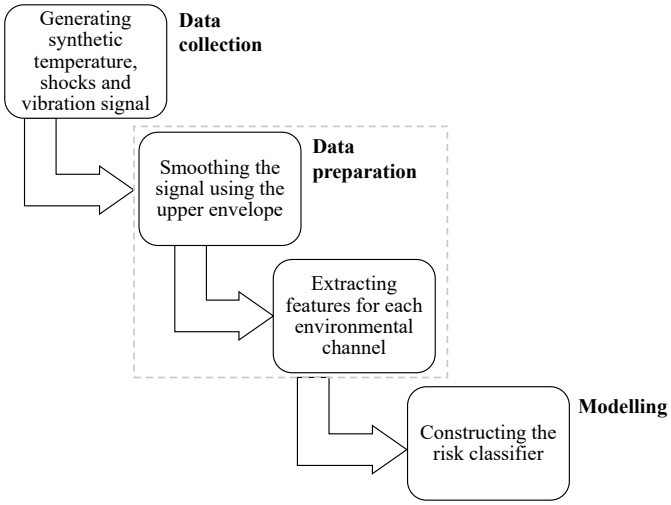


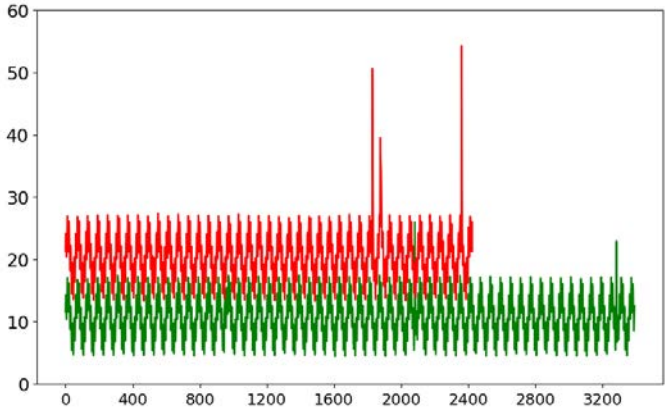
Figure 1. Proposed method.

ness, including sinusoidal waves with random time-variant amplitude and frequency, Gaussian noise, random spikes, and random shocks with random decay rates. Specifically, low-risk time series data exhibit lower parameter values for random number generation than high-risk time series data. For instance, the mean and standard deviation for generating low-risk temperature data's Gaussian noise are set to 40 and 3, respectively, while the amplitude for temperature shocks falls between 30 and 70. On the other hand, the mean and standard deviation for generating high-risk temperature data's Gaussian noise are set to 80 and 10, respectively, while the amplitude for temperature shocks falls between 50 and 100.

2.2. Data preparation

To effectively use the generated environmental data, preprocessing and optimal feature extraction are required. The preprocessing step consists of smoothing the signal using the upper envelope of the signal, as shown in Figure 2. After the preprocessing step, the environmental features can be extracted. For each environmental channel (i.e., temperature, shocks, and vibration) we compute two features based on dwell periods and two features based on load cycles, using the rainflow cycle counting method for the latter (Lee & Tjhung, 2012).

(a) Raw signal



(b) Preprocessed signal

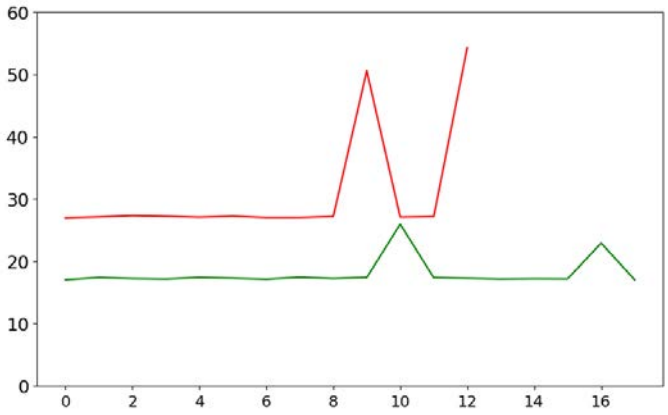


Figure 2. Generated vibration signal for a high-risk run (red) and a low-risk run(green), before and after preprocessing.

2.3. Feature extraction using rainflow cycle counting

Rainflow cycle counting is a method used in fatigue analysis to quantify the number of stress cycles experienced by a component or material (Endo, 1974).

The process involves analyzing a time series of stress or strain data to identify and count individual cycles. These cycles represent the repeated loading and unloading of a material, which can lead to fatigue failure over time. Rainflow cycle counting is especially useful for irregular or variable ampli-

tude loading conditions, where the stress levels vary over time (Lee & Tjhung, 2012).

The method consists of four steps, as illustrated in Figure 3:

1. *Hysteresis filtering* (Figure 3a) entails removing cycles smaller than an amplitude gate that contribute minimal damage. This is accomplished by setting a gate with a specific amplitude. Any cycle with an amplitude below this gate is excluded from the load-time data. The gate is projected sequentially from left to right starting from each turning point in the time series. If a turning point falls below the gate's threshold, it is omitted from the time history. (Endo, 1974)(Lee & Tjhung, 2012).
2. *Peak-valley identification* (Figure 3b) consists of locating the points in the data where the direction of the signal reverses. In a cycle, only the highest and lowest values are pertinent for fatigue life assessments. Therefore, any intermediate data points between these extremes within a cycle can be disregarded as they do not contribute to the fatigue calculation. (Endo, 1974)(Lee & Tjhung, 2012).
3. In *discretization* (Figure 3c), the amplitude dimension of the signal is divided into a set number of equal bins. Each data point is then mapped to the center of its corresponding bin to facilitate cycle counting. Centering the data samples within their bins slightly modifies their amplitudes, therefore it is crucial to utilize an adequate number of bins in the analysis to minimize significant alterations in amplitudes (Endo, 1974)(Lee & Tjhung, 2012).
4. In *four-point counting* (Figure 3d), the identified peaks and valleys are connected to form hysteresis loops, or closed paths that represent complete stress cycles (Endo, 1974)(Lee & Tjhung, 2012). This is done using the following steps:
   (a) Select four consecutive points: $S1, S2, S3, S4$.
   (b) Compute inner stress: $|S2 - S3|$.
   (c) Compute outer stress: $|S1 - S4|$.
   (d) If the inner stress range is less than or equal to the outer stress range, a cycle is counted, otherwise it is not counted.

Using the method described above, the extracted features are as follows:

1. *average cycle mean*, where the cycle mean represents the mean values of the initial and final points of a cycle
2. *average cycle range*, where the cycle range represents the absolute difference between the initial and final points of a cycle
3. *dwell time*, representing the cumulative time during which the signal oscillation is lower than a set threshold
4. *dwell damage*, representing the average amplitude during the dwell time

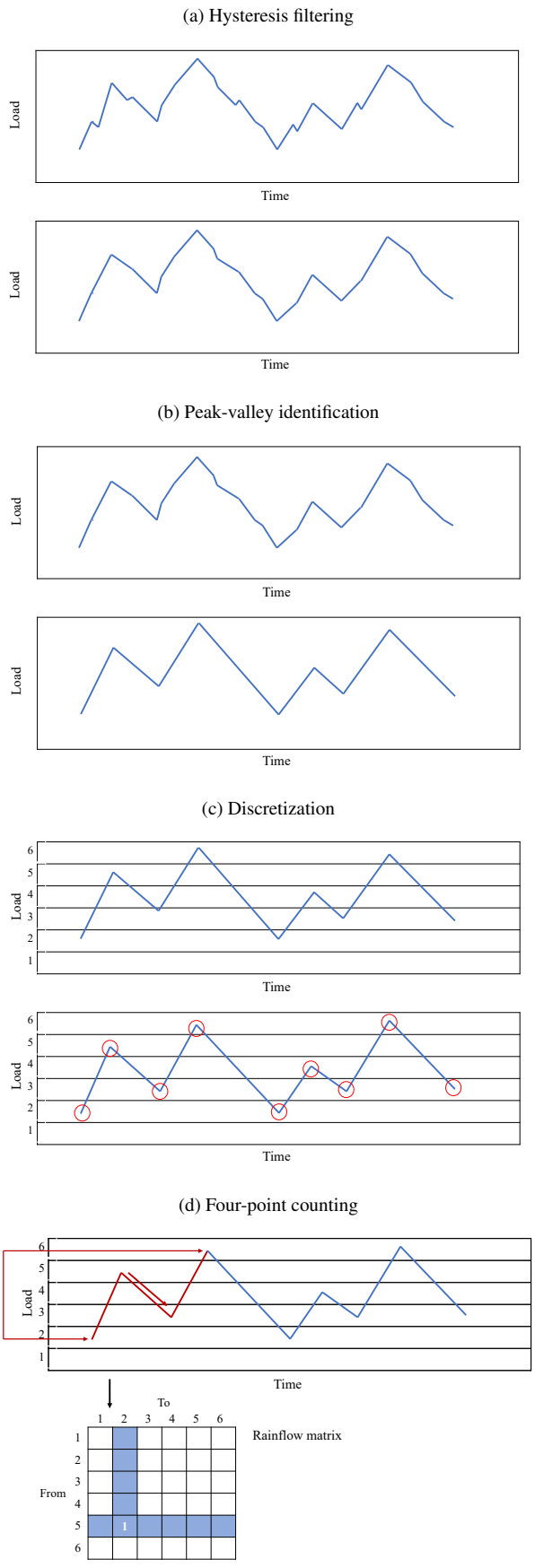


Figure 3. Rainflow cycle counting steps.

		Predicted label	
		Positive	Negative
Actual label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Figure 4. Confusion matrix for a binary classification problem.

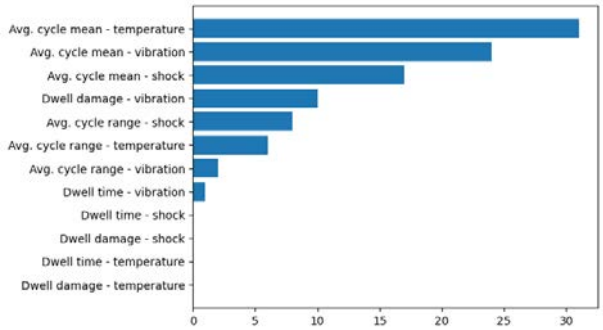


Figure 5. Model feature importance.

2.4. Modelling

We model the problem as a binary classification problem, where we interpret the positive class as high environmental risk, and the negative class as low environmental risk.

For risk classification three models were trained: logistic regression (LaValley, 2008), random forest (Biau & Scornet, 2016), and extra trees (Geurts, Ernst, & Wehenkel, 2006). The random forest and the extra trees models consist of an ensemble of 100 trees, and the Gini index was used as the splitting criterion. Logistic regression, as well as the ensemble tree-based models are less prone to overfitting and thus have the potential to generalize better to real data.

3. Case Study

A number of 1128 examples were generated, out of which 80% were used for training and 20% for testing. The training set was further split into train and validation sets in the same ratio using k-fold cross validation with 10 folds. The data were split as to preserve the class balance.

To evaluate the models on a labeled subset of the data we make use use of the confusion matrix (Fawcett, 2006), illustrated in Figure 4. In a binary classification problem, the confusion matrix has four sections:

1. True positives (TP): the number of instances where the model correctly predicts the positive class (high risk).
2. True negatives (TN): the number of instances where the model correctly predicts the negative class (low risk).
3. False positives (FP): the number of instances where the model incorrectly predicts the positive class.
4. False negatives (FN): the number of instances where the model incorrectly predicts the negative class.

To compare the models, we use the area under the receiver operating characteristic (ROC) curve (ROC AUC score). The ROC curve plots the true positive (TP) rate, defined as

$$\frac{TP}{TP + FN},$$

against the true negative (TN) rate, defined as

$$\frac{TN}{TN + FP}$$

It is a graphical representation of a binary classifier at different classification thresholds. The ROC AUC score is represented by the area under the ROC curve, where a score of 0.5 indicates a random model (Bradley, 1997).

The three trained models output a ROC AUC score of 1 on the validation set, indicated in Table 1. The application of the trained models is to assess environmental risk on electronic boards. Therefore, an important aspect is the ability of the model to successfully transfer knowledge from synthetic data to real data. In this regard, the stochastic features of the extra trees represent an advantage for increasing robustness (Geurts et al., 2006).

Table 1. Comparative ROC AUC score for the three trained models.

Model	ROC AUC score
Logistic regression	1.00
Random forest	1.00
Extra trees	1.00

We evaluate feature importance for the classification problem using Shapley values. This step helps to reduce feature redundancy and improve model interpretability. Shapley values are a method derived from cooperative game theory that has been adapted for use in explaining the predictions of machine learning models. They provide a way to fairly assess the impact of each feature for a particular prediction in a model (Merrick & Taly, 2020).

Using this method, Figure 5 indicates that for the extra trees model, the most impactful features are the average cycle means on each environmental channel, which is consistent with the feature correlation matrix in Figure 6.

Feature correlation in a machine learning model refers to the

Figure 6. Feature correlation matrix.

degree to which the variables (features) in the dataset are related to each other, as well as with the target variable. For this experiment we use the Pearson correlation coefficient (Kendall & Stuart, 1973) and we specifically study the correlation between the features and the target variable, denominated as risk. In Figure 6 we notice the highest correlation between the average cycle means on the temperature, shock and vibration channels, and the risk variable.

In the second iteration of experiments, we restrict the training to these three features.

During the validation step, the model achieves promising classification results, as indicated by the confusion matrix in Table 2. Based on the confusion matrix, we define the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1score = \frac{2 * TP}{2 * TP + FP + FN}. \tag{4}$$

The results on the validation set are consistent with the results on the test set after the training is completed, which can be seen in Table 3, despite the 0.52 score for data drift. Data drift indicates a difference in the statistical properties of the data. Therefore, the classification scores prove the robustness of the extra trees model and the potential for such a model to be used for assessing risk on real data.

Table 2. Confusion matrix on the validation set, where the positive class is equivalent to a high-risk run and the negative class is equivalent to a low-risk run.

	Predicted positive	Predicted negative
True positive	93	0
True negative	0	86

Table 3. Metrics measured on the test set.

Metric	Value
Accuracy	1.00
Precision	1.00
Recall	1.00
F1 Score	1.00
ROC AUC	1.00
Data drift	0.52

4. Conclusion and Future Work

This paper presented a data-driven approach for assessing environmental risk in electronic boards based on supervised machine learning. The method makes use of synthetic data and consists of extracting features with respect to dwell time and load cycles, showing that the latter have a larger impact on the performance of the models. The extra trees model achieves promising results on the synthetic data, but further work is needed to address the potential mismatch between training and test data in practical applications.

To address this issue, we plan to collect real-world environmental data and use it to fine-tune the model to better handle the variability of different environments. Additionally, we could explore the use of transfer learning techniques to adapt the model to new environments and improve its robustness to different types of data.

Overall, the proposed approach shows potential for assessing environmental risk in electronic boards, but further research is needed to optimize the model for real-world applications.

References

Bhargava, C., Sharma, P. K., Senthilkumar, M., Padmanaban, S., Ramachandaramurthy, V. K., Leonowicz, Z., ... Mitolo, M. (2020). Review of health prognostics and condition monitoring of electronic components. *IEEE Access*, *8*, 75163-75183. doi: 10.1109/ACCESS.2020.2989410

Bhat, D., Muench, S., & Roellig, M. (2023). Application of machine learning algorithms in prognostics and health monitoring of electronic systems: A review. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, *4*, 100166. doi: 10.1016/j.prime.2023.100166

Bhatnagar, S., Cassou, M. L., Masry, Z. A., & Mosallam, A. (2021, June). Data-Driven Fault Detection Method for Electronic Boards in Intelligent Remote Dual-Valve System. In *PHM Society European Conference* (pp. 1–7). doi: 10.36001/phme.2021.v6i1.2903

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Endo, T. (1974). Damage evaluation of metals for random or varying loading. In *Proceedings of the 1974 Symposium on Mechanical Behavior of Materials* (p. 371-380).

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861-874. Retrieved from https://www.sciencedirect.com/science/article/pii/S016786550500303X (ROC Analysis in Pattern Recognition) doi: https://doi.org/10.1016/j.patrec.2005.10.010

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Gu, J., Barker, D., & Pecht, M. (2009). Health monitoring and prognostics of electronics subject to vibration load conditions. *IEEE Sensors Journal*, *9*(11), 1479-1485.

Kale, A., Carter-Journet, K., Falgout, T., Heuermann-Kuehn, L., & Zurcher, D. (2014). A probabilistic approach for reliability and life prediction of electronics in drilling and evaluation tools. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014* (p. 519-532).

Kang, J., Varnier, C., Mosallam, A., Zerhouni, N., Youssef, F. B., & Shen, N. (2022). Risk level estimation for electronics boards in drilling and measurement tools based on the hidden Markov model. In *2022 Prognostics and Health Management Conference (PHM-2022 London)* (p. 495-500). doi: 10.1109/PHM2022-London52454.2022.00093

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics. vol. 2: Inference and: Relationsship*. Griffin.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Lee, Y.-L., & Tjhung, T. (2012). Chapter 3 - rainflow cycle counting techniques. In Y.-L. Lee, M. E. Barkey, & H.-T. Kang (Eds.), *Metal fatigue analysis handbook* (p. 89-114). Boston: Butterworth-Heinemann. doi: 10.1016/B978-0-12-385204-5.00003-3

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction* (pp. 17–38). Cham: Springer International Publishing.

Michael G. Pecht, Myeongsu Kang. (2018). *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*. John Wiley and Sons Ltd.

Mosallam, A., Kang, J., Youssef, F. B., Laval, L., & Fulton, J. (2023, May). Data-Driven Fault Diagnostics for Neutron Generator Systems in Multifunction Logging-While-Drilling Service. In *2023 Prognostics and Health Management Conference (PHM)* (pp. 171–176). doi: 10.1109/PHM58589.2023.00041

Pecht, M., & Gu, J. (2009). Physics-of-failure-based prognostics for electronic products. *Transactions of the Institute of Measurement and Control*, *31*(3-4), 309-322. doi: 10.1177/0142331208092031

Prisacaru, A., Gromala, P., Han, B., & Zhang, G. Q. (2022). Degradation estimation and prediction of electronic packages using data-driven approach. *IEEE Transactions on Industrial Electronics*, *69*(3), 2996-3006. doi:

10.1109/TIE.2021.3068681

Sobczak-Oramus, K., Mosallam, A., Basci, C., & Kang, J. (2022, June). Data-Driven Fault Detection for Transmitter in Logging-While-Drilling Tool. In *PHM Society European Conference* (Vol. 7, pp. 458–465). doi: 10.36001/phme.2022.v7i1.3362

V. Gupta, J. Kang, A. Mosallam, N. Shen, F. B. Youssef, & L. Laval. (2023, June). Automatic Fault Detection for Resistivity Systems in Logging-While-Drilling Tools. In *2023 Prognostics and Health Management Conference (PHM)* (pp. 128–132). doi: 10.1109/PHM58589.2023.00032

Vichare, N., & Pecht, M. (2009). *Method to extract parameters from in-situ monitored signals for prognostics* (No. US8521443B2).

BIOGRAPHIES

Delia-Elena Dumitru Delia-Elena Dumitru is a Data Scientist at the SLB IT center in Bucharest, Romania. She graduated in 2018 with a B.S. degree in Computer Science and completed her M.S. degree in Applied Computational Intelligence in 2020, both at the Babeș-Bolyai University in Cluj-Napoca, Romania. Her main research interests are machine learning, computer vision and PHM.

Jinlong Kang is currently a data scientist at SLB technology center in Clamart, France. He received the B.S. degree in Industrial Engineering in 2016 and the M.S. degree in Mechanical Engineering in 2019 both from University of Electronic Science and Technology of China, and the Ph.D. degree in automatics in 2024 from University of Franche-Comté in France. His main research interests are Prognostic and Health Management (PHM), maintenance decision-making, data mining and machine learning.

Ahmed Mosallam is the Data Science AI European Hub Manager at SLB technology center in Clamart, France. He has his Ph.D. degree in automatic control in the field of PHM from University of Franche-Comté in Besançon, France. His main research interests are signal processing, data mining, machine learning and PHM.

Alejandro Olid Gonzalez is currently a data scientist at SLB in Madrid, Spain. He obtained his B.Sc. degree in Physics in 2013 and his M.Sc. degree in Astrophysics in 2017. His current research interests are Prognostic and Health Management (PHM), machine learning, and simulations of physical systems.

A Flexible Methodology for Uncertainty-Quantified Monitoring of Abrasive Wear in Heavy Machinery Using Neural Networks and Phenomenology-Based Feature Engineering

Thomas Bate[1], Marcos E. Orchard[2], and Nicolás Tagle[3]

[1,2] *Department of Electric Engineering, Universidad de Chile, Santiago, Metropolitan Region, 8370451, Chile*
thomas.bate@ing.uchile.cl
morchard@u.uchile.cl

[3] *Minera Los Pelambres, Santiago, Metropolitan Region, 7550162, Chile*
ntagle@pelambres.cl

ABSTRACT

This paper introduces a cutting-edge methodology for the monitoring of abrasive wear, particularly focusing on SAG (Semi-Autogenous Grinding) mills liners. The lack of a regular inspection regime has historically led to opportunistic and thus, irregular wear measurements that are challenging to integrate into machine learning algorithms for condition-based maintenance. The study unveils a virtual sensor designed to estimate the mill liner's remaining thickness, aiming to offer daily updates and assist the maintenance team in determining the optimal timing for liner replacements without the need for halting operations. This approach is positioned as a strategic response to the critical need for efficient maintenance strategies, addressing the inherent challenges in real-world industrial settings where data quality may be poor and operational realities dominate. A significant aspect of this methodology is its emphasis on uncertainty quantification, vital for informed maintenance decision-making. This novel approach has been successfully applied to SAG mills at Minera Los Pelambres, showcasing its potential for broader applications across scenarios characterized by sporadic data collection. The results showcase an error of ±7.4254 mm of remaining thickness on the validation set, demonstrating the effectiveness of the methodology. The key contributions of this work lie in its ability to utilize low-quality data effectively and its low complexity, reducing barriers to implementing predictive health monitoring (PHM) algorithms. The successful implementation highlights the methodology's adaptability and flexibility, marking a significant advancement in the domain of maintenance strategy for the mining industry.

1. INTRODUCTION

SAG (Semi-Autogenous Grinding) mills are indispensable assets in the mining industry, serving as the cornerstone of ore processing operations. The significance of these mills cannot be overstated, as any downtime incurred due to maintenance activities can translate into substantial financial losses. The costliness of SAG mill stoppages underscores the critical need for effective maintenance strategies to ensure continuous operation and productivity.

Within the maintenance team at Minera Los Pelambres, there arose a strategic initiative aimed at reducing the duration of mill downtime attributed to inspections. To support this endeavor, the concept of the virtual remaining liner sensor was conceived. A virtual sensor, by definition, offers an approximation of a state based on other measurable variables or states, serving as an indirect measurement. In this instance, the objective was to estimate the remaining thickness in millimeters of a specific component of the mill liner online and with daily frequency updates. The overarching goal was to provide the maintenance team with a decision support tool to determine the optimal timing for liner replacement without necessitating mill shutdowns solely for inspections. One of the major challenges in implementing this system was the absence of a regular inspection schedule for the mill. Historically, inspections were performed opportunistically, aligned with planned mill shutdowns. This approach resulted in irregular wear measurements, complicating their utilization in ML algorithms designed to predict wear and determine the remaining liner thickness accurately.

The endeavor to minimize downtime due to maintenance activities has long been a focal point, with condition monitoring tasks representing approximately 13% (Kawahata, Schumacher, & Criss, 2016) of total mill downtime. Traditional

approaches to address this challenge, such as discrete element method (DEM) (Wu, Che, & Hao, 2020) simulations, have proven to be exceedingly complex and costly to implement in productive environments, necessitating expensive software and extensive time investments to achieve realistic simulations.

This paper presents a methodology specifically tailored to address the challenges inherent in real-world industry environments, where data quality is often poor, and operational considerations are paramount. By focusing solely on wear monitoring issues prevalent in the industry, this methodology emphasizes the importance of incorporating operational insights into the model design to ensure effective utilization by maintenance teams. The majority of degradation monitoring algorithms are developed using synthetic data or data obtained under suitable acquisition settings, like the measuring method proposed in (Powell & Chandramohan, 2011) (appropriate and stable sampling rates, low measurement error). However, few academic works focus on solving real-world problems where data quality is poor, as the results are naturally less impressive than those generated in studies with high-quality laboratory data, enabling the use of state-of-the-art machine learning algorithms to achieve high precision (Li et al., 2022). The most significant contribution of this work is to provide a methodology that utilizes low-quality data to its fullest potential. The low complexity of the method reduces the barriers currently faced by the industry in implementing PHM algorithms.

This challenge is predominantly practical rather than theoretical, as the methodology was conceived with real-world industrial scenarios in mind, the approach is quite easy to implement. Leveraging neural networks and feature engineering based on phenomenology, the proposed approach is successfully implemented to monitor the liners of SAG Mills at Minera Los Pelambres, providing a decision-support tool for the maintenance team. The methodology is designed to infer deviations from an average wear rate curve, utilizing features that represent stress factors on the mill derived from both historical lining data and real-time mill operation information. The predictive modeling aspect of the methodology employs neural networks, these networks offer accurate inferences of wear progression, allowing for proactive maintenance strategies and the timely identification of potential issues. Complementing the data-driven approach, the methodology incorporates feature engineering grounded in the phenomenology of abrasive wear. This ensures that the monitoring scheme is not solely reliant on learned patterns but also integrates domain knowledge, enhancing interpretability and generalization. A distinctive feature of the methodology is its focus on uncertainty quantification in wear monitoring during the online operation of the model, this is crucial for decision-making, providing the maintenance team with insights into the reliability of wear assessments and facilitating the prioritization of maintenance interventions.

Incorporating uncertainty quantification in industrial monitoring is essential for enhancing intelligent maintenance decision-making. This approach provides a probabilistic perspective on operational data, facilitating a more nuanced understanding of equipment behavior and maintenance needs. The key benefits include:

1. **Enhanced Decision-making**: Uncertainty quantification allows for informed, risk-aware decision-making. By understanding the range of possible outcomes and their probabilities, maintenance teams can make decisions that improve safety, operational efficiency, and financial performance.

2. **Optimized Maintenance Scheduling**: It aids in identifying the most opportune moments for maintenance actions, balancing preventive and corrective strategies. This optimization minimizes operational disruptions and costs while extending equipment lifespan.

3. **Risk Management**: Understanding model uncertainty helps in managing the risks associated with maintenance activities. Identifying high-risk scenarios enables prioritization of critical maintenance interventions, ensuring operational continuity and safety.

4. **Confidence in Predictive Models**: Quantifying uncertainties builds confidence in predictive maintenance models by transparently communicating their reliability. This transparency is crucial for trust among operational staff and stakeholders.

The successful implementation on SAG Mill liners at one of the world's largest copper mine validates the methodology's efficacy. Beyond its application to this specific context, the methodology is highlighted for its flexibility and adaptability to other scenarios. Notably, it accommodates few and irregular measurements of the asset's state, making it applicable in situations where data collection may be limited or sporadic.

However, unlike traditional approaches such as discrete element method (DEM) simulations which are complex and costly to implement, this methodology provides a straightforward and cost-effective alternative. By leveraging low-quality data and emphasizing uncertainty quantification, we address the practical challenges faced in real-world industrial settings. This approach not only simplifies the implementation of PHM algorithms but also ensures robust predictions even with sporadic data collection. Additionally, the incorporation of phenomenology-based feature engineering enhances the interpretability and reliability of the model, setting it apart from purely data-driven methods.

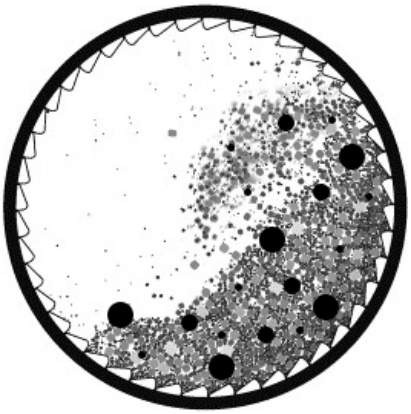


Figure 1. SAG Mill illustration.

2. SAG MILLS BACKGROUND

SAG mills consist of rotating drums containing metal balls that cascade and impact against the mineral (Figure 1), effectively grinding it. The collision between grinding balls and mineral particles fractures and further grinds the mineral, producing finer material. SAG mills are distinguished by their large diameter and short length compared to ball mills. The interior of the mill is lined with lifter plates to lift the material inside, facilitating the cascading material flow for grinding. Figure 1 illustrates this operation. Various lifter configurations and positions line the interior of the mill. Of particular interest is the discharge end cap, where worn or fractured components could lead to ball escape and downstream processing issues.

To assess the condition of the mill liners, a procedure known as a ”faro” is typically conducted. A faro is akin to a radiographic examination of the mill, providing precise measurements of the liner condition. However, these stoppages are costly. There has never been a consistent schedule for the faros, therefore the sampling rate is inconsistent, the only constant measurement made is at the end of the liners life when it is removed from the mill. The virtual sensor developed in this study aims to reduce the frequency of faro inspections, offering online monitoring capabilities to track the remaining liner thickness, particularly focusing on the discharge end cap, where liner failure poses significant operational risks.

The availability of SAG mills is paramount in mineral processing, as every hour of downtime translates to substantial financial losses, valued in thousands of dollars. Optimization of maintenance activities is crucial to minimize mill stoppages, balancing the risk of failure with maintenance requirements.

Given the criticality of mill uptime, any condition monitoring initiative aiding in optimal maintenance scheduling adds significant value. The virtual sensor developed in this study aligns with this objective, providing the maintenance team with decision support tools to optimize faro inspection schedules without necessitating mill shutdowns. This approach not only reduces downtime but also mitigates the risk of operational disruptions downstream.

The current market offers various solutions for monitoring the liners of SAG mills, each with its own set of advantages and challenges. Many of these solutions rely on expensive hardware or require interventions directly on the mill cylinder (Dandotiya, Lundberg, & Wijaya, 2011). However, few effectively leverage historical data to optimize monitoring processes.

One significant challenge is the difficulty of integrating additional monitoring equipment into large-scale machinery already in production. SAG mills are massive industrial units critical to the mineral processing chain, and any modifications or additions to these machines must be carefully implemented to avoid disrupting operations. Installing new monitoring devices often involves intricate engineering work and may require halting production for extended periods, leading to significant downtime and revenue loss for mining companies.

Moreover, the harsh operating conditions within SAG mills present further challenges. These mills operate in environments characterized by high temperatures, dust, and vibrations, which can adversely affect the performance and longevity of monitoring equipment. Ensuring the reliability and durability of monitoring devices under such conditions is essential but often requires additional investments in ruggedized hardware and protective enclosures. Mill liners are located deep within the mill cylinder, necessitating specialized equipment and skilled personnel for installation and maintenance tasks. Any monitoring solution that requires frequent access to the liners may incur significant logistical challenges and operational disruptions.

In light of these difficulties, there is a growing need for innovative monitoring solutions that can leverage existing data infrastructure and minimize disruptions to mill operations. Solutions that harness historical data and employ non-intrusive monitoring techniques offer promise in this regard, providing valuable insights into liner wear patterns while minimizing the need for costly hardware installations and production stoppages.

At Minera Los Pelambres, where three SAG mills—SAG1, SAG2, and SAG3—are operational, the focus is primarily on SAG1 and SAG2 due to their comprehensive data records and identical machinery specifications. The successful replication of the methodology for SAG3 underscores its potential scalability and applicability across multiple mill units, albeit not covered in this document.

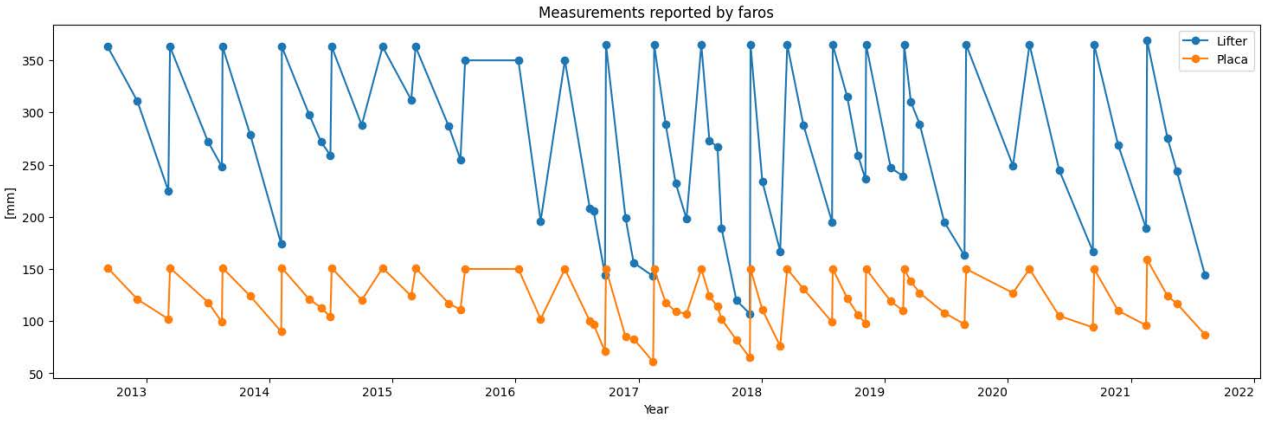


Figure 2. Historic remaining liner mesurements SAG1.

3. Data and Processing

The methodology involves two primary data sources: operational tags related to the mill's functioning and mineralogy, and liner wear measurements reported through inspections. Operational tags, which are time series data from sensors or states, can be extracted at various granularities, they are stored via Pi Systems, a platform for operational data management developed by OSIsoft. It plays a crucial role in this project by serving as the source of operational data. This platform captures, stores, analyzes, and visualizes real-time data from industrial processes. Pi System has its own data cleaning protocols which are critical to ensuring data quality and reliability before analysis, and these protocols were considered in the project's data processing strategy. Given that Pi Systems are commonly used across various industries to manage data, it is important to delve into how data is handled and processed within the methodology. The first step in data processing involves cleaning and imputing these tags. The main issues addressed in the data are outliers, non-numeric values, and missing values. The cleaning process categorizes tags into four types:

- Tags representing percentages: Values above 100 or below 0 are set as NaN (empty value).
- Tags for positive variables with distributions similar to normal: Values below zero and those above the 99th percentile distribution (outliers) are set as NaN.
- Binary variable tags, which include two variables:
  - Rotation direction: Non-numeric values are present, with two relevant states indicating clockwise and counterclockwise rotation. Clockwise is replaced with 1, counterclockwise with -1, and other messages with NaN, to then interpolate them with the closest value.
  - Mill state: Relevant messages indicate whether the mill is stopped (0) or operating (1). Other messages are set to 0.
- Tags that do not require cleaning.

Non-numeric values, often error messages from the tag storage system, are addressed next. Messages indicating a value above/below defined ranges are replaced with the tag's post-cleaning maximum/minimum value. Remaining NaN values are imputed linearly. The tags are cleaned on an hourly basis, including:

- Grinding hardness
- Feed water
- Load cell
- Stator current
- Noise detector
- F80
- Rotation
- Granulometry (of the incoming mineral) 100, 200, 325, 48, 65, 125 Inches
- Filling level
- Solids percentage
- Power
- Pressure
- Noise
- Speed

Following the removal of non-numeric values, data transformation begins, incorporating liner wear measurements. The term 'campaign' refers to the lifespan of the lining. A campaign begins when the lining is installed and ends when it is retired. These measurements are taken with the mill stopped and vary across campaigns, with up to 5 intermediate measurements in some campaigns and a median of 3. The above Figure (2) shows SAG1 grate wear monitoring over time, with two curves representing different grate positions as reported by faros. Points on the curves are measurements, and lines

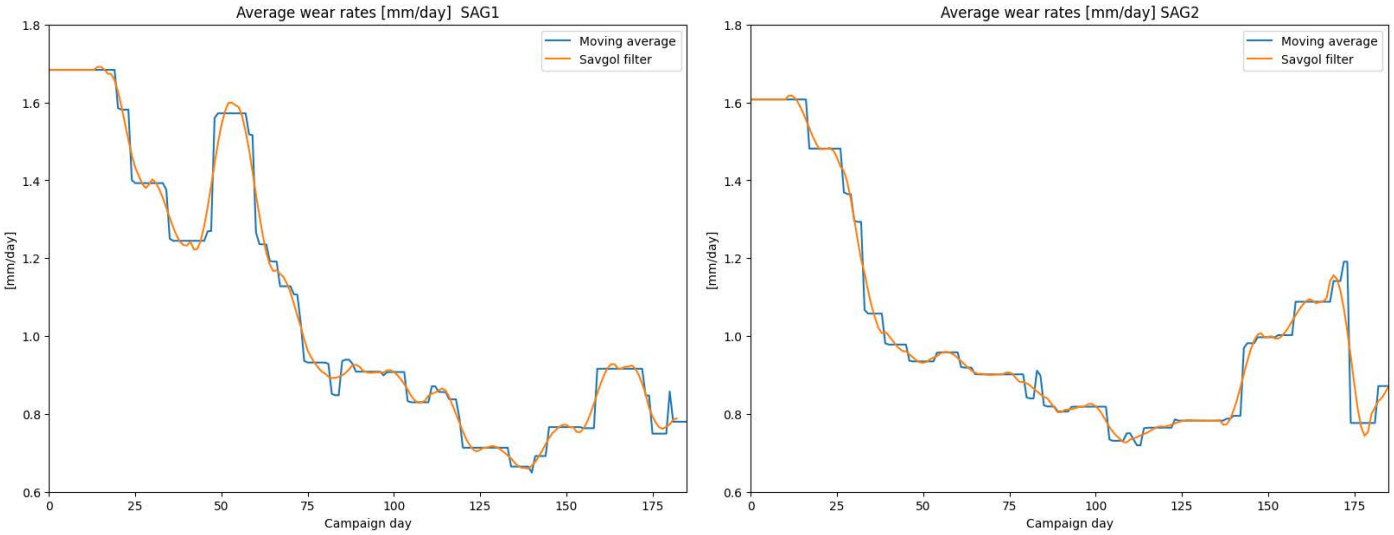


Figure 3. Average wear curves for the lifter position.

represent linear interpolations between points. This highlights the irregularity and varying wear rates in SAG1's history, a pattern consistent across other mills. Almost all the campaigns have a measurement at the start of the lining's lifespan and one near to its end.

Within each campaign, segments between contiguous measurements are defined to calculate consumed thickness and total operational hours, yielding a wear rate in [mm/hour], which is converted to [mm/day]. The objective is to update the liner's remaining millimeters daily. To achieve this, average wear rate curves for SAG1 and SAG2 are calculated (Figure 3). Operational days within each campaign are assigned a wear rate, and using all campaigns, daily wear rates are averaged to produce the curves shown in the next figure. A Savgol filter is applied to obtain the final average wear curve used throughout the study.

The difference between the curves is mainly due to SAG2 receiving more recirculated material, which is less abrasive. Finally, new variables are generated, including accumulated minaral-flows, moving averages, time window dispersions, and others detailed below, aggregated daily and indicating the mill's operational percentage per day.

- Accumulated mineral-flow.
- Clockwise mineral-flow.
- Counter-clockwise mineral-flow.
- Accumulated counter-clockwise flow.
- Accumulated clockwise flow.
- Velocity dispersion over a 72-hour window.
- Accumulated velocity dispersion sum.
- Accumulated power (electric consumption).
- Load cell weight moving average.
- Load cell weight dispersion over a 72-hour window.
- Accumulated load cell weight dispersion sum.
- Noise power moving average.
- Noise power dispersion over a 72-hour window.
- Accumulated noise power dispersion sum.
- Operational day of the campaign.
- Day of the campaign.
- Percentage of clockwise operation time during the campaign.

These enhancements prepare the variables for model input, with accumulations specific to each campaign.

4. Proposed Methodology

The initial decision was to utilize a unified model for both SAG1 and SAG2, justified by the fact that they are essentially the same machinery, albeit with some operational differences. All input variables are aggregated on a daily level, aiming for the model to approximate daily wear of the mill, subsequently accumulating it throughout the campaign for real-time wear monitoring.

After defining the model's input variables, the next step was addressing its output. Resulting from the daily aggregation of data, each data row contains the daily average of tags and created variables, a daily mill utilization percentage, and the interpolated wear rate for that specific day. However, the model does not output this daily wear rate directly; instead, it uses the deviation from the interpolated wear rate compared to the previously mentioned average rates. The creation of the output for the regresor and the histogram of this deviations is

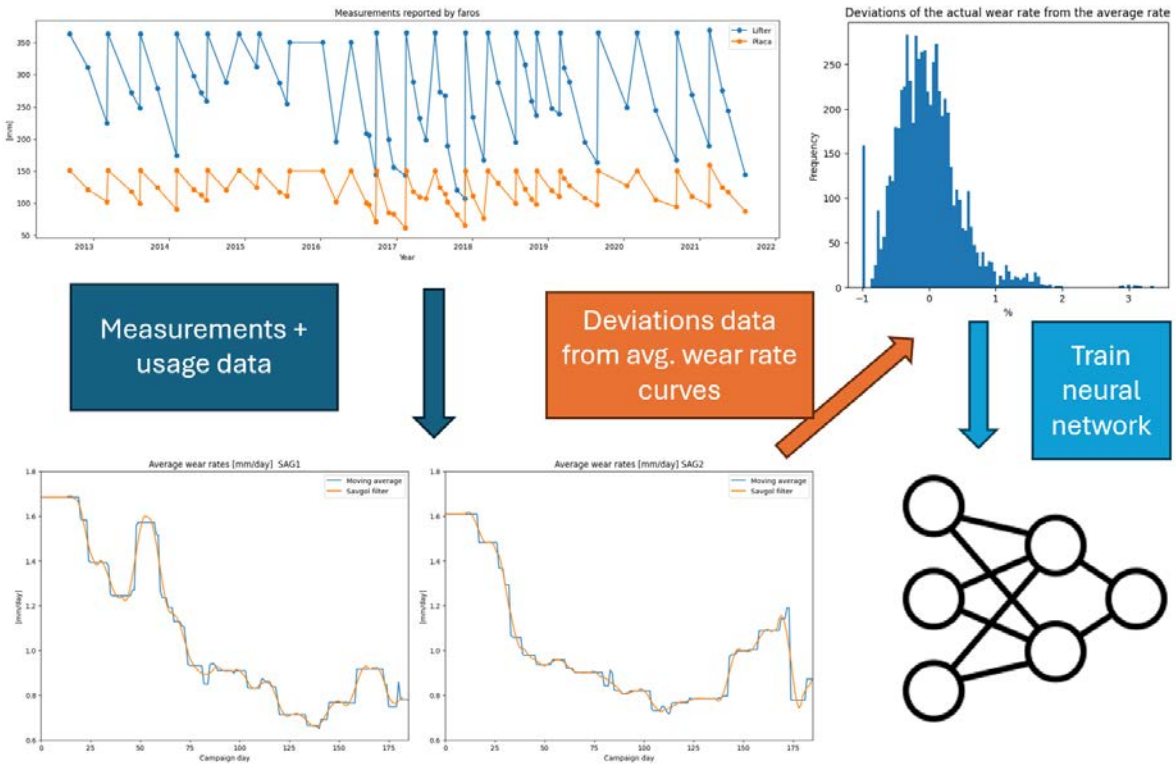


Figure 4. Computation of the deviations from the average wear rate curve.

illustrated in Figure (4). The wear expression in Equation (1) is derived based on the cumulative effect of daily operational conditions on the liner's wear rate. Specifically, we start with the basic principle that the wear on any given day is influenced by both the average wear rate for that day and deviations due to specific operational conditions. Mathematically, this can be expressed as:

$$s(T) = \sum_{i=1}^{T} (\delta_i + \delta_{model}(x_i)) * \alpha_i \tag{1}$$

where T denotes the current day, s represents the lining state measured in millimeters, δ_i is the average wear rate for operational day i, δ_{model} is the model's output, x_i is the model input, and α_i is the day i utilization percentage.

The training set was determined by selecting campaigns with a significant number of measurements to create average wear curves. Considering the irregularity of measurements per campaign and their impact on model training, campaigns with only one intermediate measurement, typically towards the campaign's end, were excluded from the training set (and included in the validation set) due to their limited information contribution about the wear pattern. Thus, the training set comprised 36 campaigns, with a testing set of 8 campaigns chosen for their recency (in order to test the methodology with the most recent campaigns) and lack of more than one measurement.

To train the model, only days from the training campaigns with at least 21 operational hours were used, addressing the distinct data distribution during mill stoppages or startups. The model (in production and testing) was fed all days regardless of operational hours, adjusting outputs by the calculated utilization percentage to prevent unexpected results from low-operation days.

Several types of regressors were tested, including linear regression, decision trees, and support vector machines; however, the neural network was chosen due to its superior interpolation capability and its effectiveness in handling the complexities and variations present in the historical wear data. A Multi-Layer Perceptron (MLP) neural network with three hidden layers and slight dropout was trained using the training set, aiming to minimize the error between the prediction and the actual deviation from the average wear rate curve. The network's performance was then tested against the validation campaigns, focusing on minimizing the projection error, defined as:

$$e_{proy} = \sqrt{\sum_{k=0}^{N} (s^k(T) - s^k_{real}(T))^2} \tag{2}$$

where k indexes the validation campaigns, and $s^k_{real}(T)$ rep-

resents the actual liner measurements for that campaign in millimeters.

The network's output represents a percentage deviation, which is then converted to millimeters of liner wear before being adjusted by the daily utilization percentage, ensuring the model accurately reflects operational impact on wear. The distinction between the metric training the neural network and the projection error metric highlights the goal of accurately modeling the actual mill state through the accumulation of neural network results.

5. RESULTS

Recalling from the previous section, the model was developed for the component known as the grate, which, as shown in Figure 2, is monitored at two positions on the grate, named lifter and plate. Therefore, there are two models, one for each position, and the results for both models, which follow the exact procedures described earlier, will be reported. The best model generated for the plate achieved a projection error of 7.4254 mm, whereas the lifter model had an error of 8.701 mm. Below is the table highlighting the projection errors for both models, two example campaigns (Figures 5 and 6) are given in order to illustrate the performance of both models, comparing their results with the faros and with the curve generated by integrating the average wear rates (also weighted by the daily utilization rate), which will serve as a reference. The projection error obtained for those two validation campaigns is also reported.

Table 1. Reported Projection Errors

Model Position	Projection Error (mm)
Plate	7.4254
Lifter	8.701

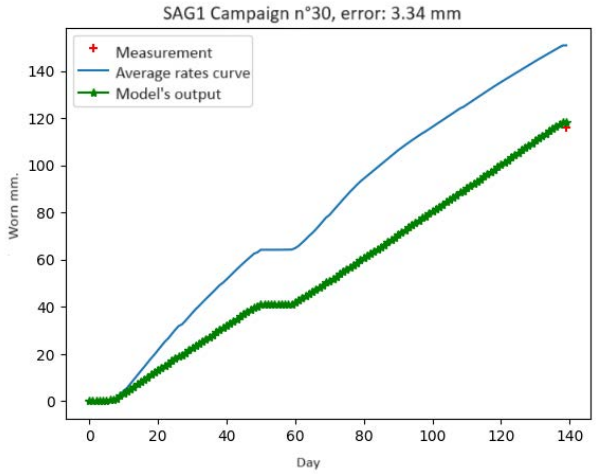


Figure 5. Model evaluation example with campaign 30 (validation) SAG.

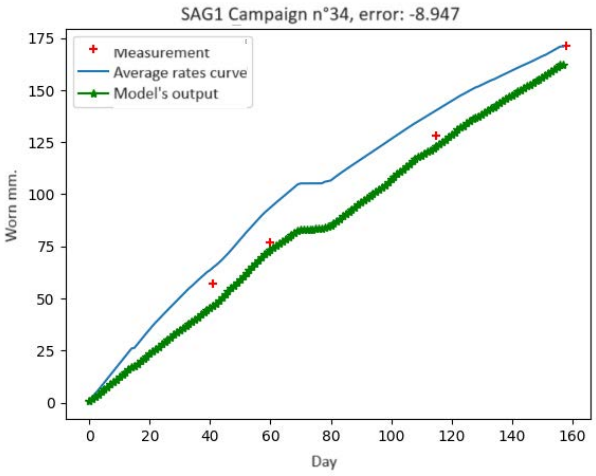


Figure 6. Model evaluation example with campaign 34 (validation) SAG.

5.1. Analysis

The study demonstrates the viability and effectiveness of modeling SAG1 and SAG2 operations jointly. Given that they are identical machines whose variables operate within the same ranges despite differences in their operational patterns, a unified model approach fosters a more robust solution. This robustness stems from training a neural network with data from both mills, offering a larger dataset per model than would be available if two separate models were trained for each mill, also avoiding over-fitting on a single mill's typical operation. This approach not only improves the model's accuracy but also its general applicability across identical machinery.

A significant insight from this work is the advantage of establishing an 'average' operational reference for each asset, as exemplified by the average wear curve against which each mill's wear is calculated. This methodology allows for the development of a model that calculates deviations from 'normal' operation, ensuring predictions remain within reasonable bounds. The histogram of deviations (used as output of the regresor) showed in Figure 4 confirms that there are no significant deviations from the average, highlighting the model's reliability in providing plausible calculations from an operational perspective. This trait is critical for maintaining operator trust in the model, a confidence that could be undermined by implausible model outputs.

The project faced considerable challenges due to the scarcity and poor quality of intermediate wear measurement data. Effective data handling and processing were crucial for maximizing the utility of the available information. Campaigns with only one measurement were primarily useful for validating the model's wear projections and offered limited value for training purposes.

The obtained results are satisfactory, yet they lack a crucial aspect to function online: integrating recent measurements to adjust predictions and quantify uncertainty.

5.2. Online Operation and Uncertainty Quantification

Implementing real-time functionality and accounting for measurement updates in the model necessitates a dynamic approach to incorporate new measurements from inspections, a crucial enhancement given the model's reliance on up-to-date information. A particle filter, a key algorithm in this study, offers an effective tool for state estimation in online settings when incorporating real-time measurements. This Bayesian recursive estimator employs discrete particles to approximate the posterior distribution of the estimated state, making it suitable for online state estimation with available measurements and a system model correlating model states with measurements. It involves initialization, prediction, and correction steps, recursively calculating state estimates.

In the context of this work, a simplified particle filter was implemented as follows:

1. **Initialization**: With an initial measurement always available, particles are sampled from a normal distribution centered on this initial measurement, with variance related to measurement error. Each particle's weight is initialized as $\frac{1}{N}$, where N is the number of particles.

2. **Model Prediction**: Particles follow the model's trajectory, with added noise to introduce variability among the particles.

3. **Measurement Update**: Upon receiving a measurement, the posterior state distribution is calculated using the particles, with new weights computed based on each particle's likelihood given the measurement. If a weight disparity condition is triggered, a resampling step occurs. The process returns to the previous step upon completion.

This particle-based approach generates a probability distribution of the state to be estimated, acknowledging and addressing the inherent uncertainty, thus offering a solution that manages the uncertainty associated with the state estimation process effectively. The variance of the particles in the particle filter was calculated based on the projection error, allowing the filter to produce calibrated uncertainty quantification.

For instance, in campaign 34 of SAG1 (Figure 7), the particle filter had minimal impact due to the model's consistent accuracy. However, in campaign 27 (Figure 8) of SAG2, a significant deviation was corrected by the filter upon the third measurement, thereby improving model performance towards the campaign's end.

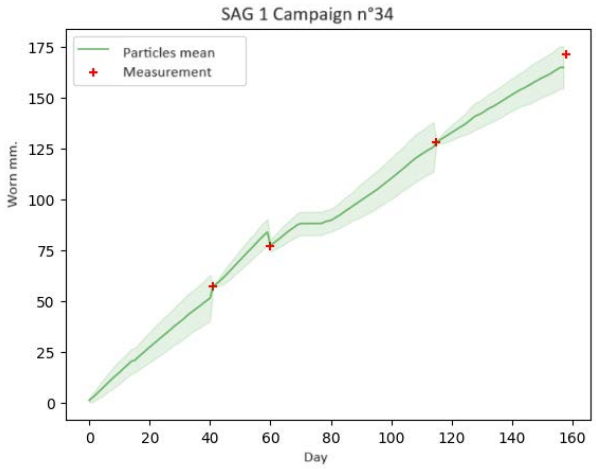


Figure 7. On-line operation of the model with the particle filter, SAG1 campaign 34.

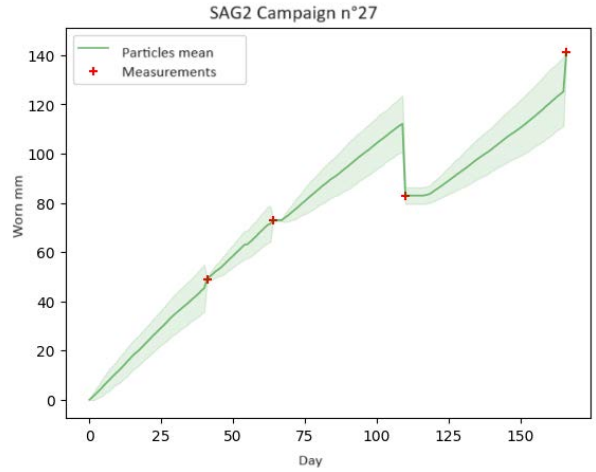


Figure 8. On-line operation of the model with the particle filter, SAG2 campaign 27

6. Conclusions

This paper presents an innovative methodology for abrasive wear monitoring in SAG (Semi-Autogenous Grinding) mills, addressing the challenge of irregular wear measurements due to the lack of a regular inspection regime. The introduction of a virtual sensor aims to estimate the liner's remaining thickness, providing daily updates to assist the maintenance team in scheduling liner replacements efficiently. This method proves critical in enhancing maintenance strategies, particularly in environments where data quality may be compromised and operational realities prevail. A key feature of this approach is the emphasis on uncertainty quantification, which is crucial for informed maintenance decision-making.

The successful application of this methodology to SAG mills at Minera Los Pelambres demonstrates its effectiveness and

potential for broader adoption. Achieving an error of ±7.4254 mm of remaining thickness for the plate position and ±8.701 for the lifter in the validation set underscores the models precision. The methodology's ability to utilize low-quality data and its simplicity are among its most valuable contributions, reducing the barriers to implementing predictive health monitoring (PHM) algorithms and marking a significant advancement in maintenance strategies for the mining industry.

ACKNOWLEDGMENT

This work has been partially supported by FONDECYT Chile Grant Nr. 1210031, and the Advanced Center for Electrical and Electronic Engineering, AC3E, Basal Project FB0008, ANID.

REFERENCES

Dandotiya, R., Lundberg, J., & Wijaya, A. R. (2011). Evaluation of abrasive wear measurement devices of mill liners.. Retrieved from https://api.semanticscholar.org/CorpusID:15963

Kawahata, K., Schumacher, P., & Criss, K. (2016, 07). Large-scale mine production scheduling optimisation with mill blending constraints at newmont's twin creeks operation:. *Mining Technology*, *125*, 1-5. doi: 10.1080/14749009.2016.1212510

Li, K., Chen, M., Lin, Y., Li, Z., Jia, X., & Li, B. (2022). A novel adversarial domain adaptation transfer learning method for tool wear state prediction. *Knowl. Based Syst.*, *254*, 109537.

Powell, M., & Chandramohan, R. (2011, 01). A structured approach to modelling sag mill liner wear - monitoring wear.

Wu, W., Che, H., & Hao, Q. (2020, 11). Research on non-uniform wear of liner in sag mill. *Processes*, *8*, 1543. doi: 10.3390/pr8121543

A Gear Health Indicator Based on f-AnoGAN

Hao Wen[1,2], Djordy Van Maele[3,4], Jean Carlos Poletto[3,4,5], Patrick De Baets[3,4] and Konstantinos Gryllias[1,2]

[1] *KU Leuven, Department of Mechanical Engineering, Celestijnenlaan 300, 3001, Leuven, Belgium*
hao.wen@kuleuven.be
konstantinos.gryllias@kuleuven.be

[2] *Flanders Make @ KU Leuven, Leuven, Belgium*

[3] *Ghent University, Soete Laboratory, Technologiepark Zwijnaarde 46, 9052 Zwijnaarde, Belgium*

[4] *Flanders Make @ UGent - Core Lab MIRO, Ghent, Belgium*

[5] *Federal University of Rio Grande do Sul, Laboratory of Tribology, Osvaldo Aranha, 99, 90035-190, Porte Alegre, Brazil*

ABSTRACT

The development of high-quality health indicators based on Artificial Intelligence (AI) for condition monitoring, reflecting the degradation process and trend, remains a key area of research. Unsupervised deep learning methods, such as deep autoencoders and variational autoencoders, are often employed to establish health indicators for rotating machinery. However, commonly used methods frequently face challenges in controlling and evaluating the quality of learned features that represent this distribution, which subsequently impacts the accuracy of the test data analysis and anomaly detection. Additionally, the empirical nature of threshold setting adds an element of uncertainty to detections.

The research propose a novel approach for constructing gear health indicators and performing anomaly detection using Generative Adversarial Networks (GAN), with a particular emphasis on the f-AnoGAN structure. The research focuses on modeling the distribution of vibration signals acquired from healthy systems using adversarial learning. By comparing test samples against this modeled distribution, the degree of similarity or dissimilarity acts as an indicator of anomalies. Owing to the generative process of the GAN architecture (creating data from randomly sampled low-dimensional noise), GAN-based modeling overcomes the limitation of autoencoders by aiming to reconstruct the continuous distribution of systems in healthy conditions from a limited set of healthy (training) samples. In this way, it offers more generalizability than traditional model learning. Moreover, this study proposes a new method for establishing thresholds based on distribution fitting by the anomaly score of healthy data. The proposed f-AnoGAN-based model and thresholding technique is applied, tested and evaluated in a gear-pitting degradation dataset and result in more accurate and timely fault detection, markedly enhancing the ability to identify subtle faults in systems over traditional methods.

1. INTRODUCTION

Gears are an indispensable element of rotating machinery, widely employed across industry, including aerospace, rail transport, and industrial sectors (Chen, Jiang, Ding, & Huang, 2022; Salameh, Cauet, Etien, Sakout, & Rambault, 2018). The malfunctioning of gears constitutes a prevalent reason for the failure of machine systems, which can result in substantial economic losses and may even pose risks to human safety (Lee et al., 2014). Consequently, monitoring gear conditions and accurately predicting component failure and fault progression are crucial.

The employment of vibration based condition monitoring at both system and component levels represents a universally endorsed technique within the realm of health monitoring for rotating machinery (Elasha et al., 2014; Teng, Wang, Zhang, Liu, & Ding, 2014; Öztürk, Sabuncu, & Yesilyurt, 2008). The meticulous measurement and subsequent analysis of vibration signals are instrumental in the precise identification of in-

cipient faults, thereby enabling the implementation of preventative and predictive maintenance prior to deterioration and corresponding issues. This proactive approach significantly contributes to the sustenance of system reliability and safety. Moreover, vibration analysis serves as an invaluable source of insight regarding the mechanical condition, since deviations in pivotal rotating elements, such as gears, manifest within the vibration signals (Zhu, Mousmoulis, & Gryllias, 2023; Hendriks, Dumond, & Knox, 2022). The utilization of signal processing tools for the examination of vibratory data aids in the extraction of critical information and indicators spanning both frequency and time-frequency domains. Nevertheless, it is imperative to acknowledge that the interpretations derived from these signal processing outcomes frequently require the expertise of seasoned operators.

With the advancement of artificial intelligence, its application in gear fault detection has gained increasing attention. Artificial intelligence, especially machine learning and deep learning methods, can process and analyze vast amounts of data, uncovering complex patterns and relationships that may be elusive to human experts. This reduces reliance on deep expert knowledge while enhancing the efficiency and accuracy of fault detection, enabling even non-experts to effectively diagnose faults. Among the various techniques, Convolutional Neural Networks (CNN) have demonstrated their versatility in state monitoring applications, including the detection and diagnosis of gear pitting faults (Zhang, Liu, Wang, & Gu, 2022; Xiang, Yang, Hu, Su, & Wang, 2022; Shi et al., 2022; Kim, Na, & Youn, 2022).

Viewing fault detection as a classification problem is a widely adopted strategy. However, obtaining clean, ample, and balanced healthy and especially faulty data, is challenging. Thus, various unsupervised one-class classification methods have been introduced. Unsupervised training methods, which infer based solely on information from healthy data, are limited by their output being the probability of a sample being normal. Thus for detection in a continuous progress, such as a degradation, this type of methods lacks of ability to represent trend. These methods primarily involve two steps: firstly, through the neural network's learning, mastering the distribution of healthy data and gauging the deviation of test data from this baseline; secondly, establishing reasonable thresholds for anomaly detection. A popular method is Deep Support Vector Data Description (DSVDD) (Ruff et al., 2018; Liu & Gryllias, 2020; Peng, Liu, Desmet, & Gryllias, 2023), which uses the Euclidean distance between hidden layer feature representations to characterize the extent of faults, allowing for trend assessment. However, DSVDD faces limitations in feature space representation capability and a lack of control over hidden layers/features.

An alternative unsupervised learning approach involves self-supervised schemes like Autoencoder (AE). By encoding and decoding complete data through neural networks, these models learn the intrinsic structure of the data (C. Zhou & Paffenroth, 2017; Ren, Sun, Cui, & Zhang, 2018; Mao, Feng, Liu, Zhang, & Liang, 2021). Yet, the characteristic of data compression in autoencoders limits their generalization ability, showing a significant dependency on the training data.

Recent years have seen generative models emerge as a new research focus. From the perspective of mechanical fault detection, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been applied primarily in data augmentation tasks (He, Tian, & Zuo, 2022; K. Zhou, Diehl, & Tang, 2023; Qin, Wang, & Xi, 2022; Wang et al., 2019). (Ding, Ma, Ma, Wang, & Lu, 2019) proposed a GAN-based anomaly detection method for bearing fault diagnosis, where the discriminator is used as an anomaly detector. (Dai, Wang, Huang, Shi, & Zhu, 2020) introduces an adversarial learning strategy to optimise the training of autoencoder(the method is also known as adversarial autoencoder) for the establishment of rotating machinery health indicators.

However, the essence of GANs lies in their use of adversarial learning to better fit the distribution of training data, allowing the direct generation of new data from this distribution. This aligns with the upstream task of various anomaly detection algorithms, which is to simulate the distribution of training data.

This study explores the potential of Generative Adversarial Networks in the task of anomaly detection for rotating machinery, based on vibration signals. It proposes a scheme for constructing a gear health indicator using GANs, along with a corresponding threshold setting and an anomaly detection system, aimed at detecting pitting initiation as early as possible. The methodology is validated on a dataset from a gear-accelerated degradation test.

The rest of the paper is organised as follows. In Section 2, the proposed anomaly detection methodology including model training, construction of health indices, and the threshold setting scheme is presented in detail. Then in Section 3 the experimental set up is described, the proposed methodology is applied on the experimental dataset and its effectiveness is analysed. The paper closes at the final section with some conclusions and the potentials of the proposed method in the field of rotating machinery health monitoring.

2. Proposed Method

The proposed detection scheme can be divided into three independent steps:

1. **Offline Distribution Learning**, by generative adversarial learning. In this step, the model is trained only by a limited number of partitioned healthy signals. The generator uses low-dimensional random noise as input and upscales it to the same dimension as the actual samples.

The objective of the generator is to produce signals from the random noise (i.e., feature space) that are as realistic as possible. This process can be considered as the model's grasp of the intrinsic structure of the training signals.

2. **Health Indicator Formation**, by the fast AnoGAN (f-AnoGAN) structure. The well-trained generator from step 1 can upscale any arbitrary low-dimensional feature set to obtain sufficiently realistic signals. This result can be interpreted as having obtained a continuous, infinite set of training samples. Therefore, for a test sample, its state related with health can be determined by whether an identical sample (or, as similar as possible) can be found within this continuous healthy set. The process of finding the corresponding sample, according to the f-AnoGAN structure, is assisted by an independently trained encoder working alongside the generator. After obtaining the corresponding sample for the test signal, the Euclidean distance is measured between signals to gauge the test sample.
3. **Fault Detection**, by a thresholding method. The discrepancy measured as outlined in step 2 is compared against a pre-determined threshold. Samples exceeding this threshold are flagged as potential anomalies, indicating a departure from the healthy signal distribution and hence, identifying possible faults.

2.1. Generative Adversarial Network (GAN) Training

2.1.1. Training Strategy

The training of GAN, depicted in Figure 1, alternates between updating the discriminator (also referred to as the Critic in the context of WGANs) and the generator. The discriminator (Critic model)'s task is to evaluate the realism of both real and generated samples, while the generator aims to produce data that are indistinguishable from real data. The key innovation of WGAN-GP (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017) lies in the gradient penalty term, which enforces a soft version of the Lipschitz constraint by penalizing the gradient norm of the Critic's output with respect to its input.

2.1.2. Loss Composition

Generator Loss: The generator's objective is to minimize the negative average score of the generated samples evaluated by the discriminator:

$$L_G(\theta_G) = -\mathbb{E}_{\tilde{x}\sim P_g}[C(\tilde{x})] \tag{1}$$

where θ_G represents the generator's parameters. The generator is trained to produce samples $\tilde{x}$ that maximize the discriminator's (critic's) score $C(\tilde{x})$, pushing it towards generating more realistic samples.

Discriminator Loss: It includes two components - the average score for the real samples and the average score for the generated (fake) samples.

The objective of the GAN's training can be expressed as:

$$\min_{\theta_G} \max_{\theta_C \in \mathcal{C}} \mathbb{E}_{x\sim P_r}[C(x)] - \mathbb{E}_{\tilde{x}\sim P_g}[C(\tilde{x})] \tag{2}$$

where θ_C represents the critic's parameters. The goal is to train the critic to assign higher scores to real samples $x \sim P_r$ and lower scores to generated samples $\tilde{x} \sim P_g$.

However, the optimizing objective (2) is still not effective enough in the practice of GAN training, and researchers are often plagued by pattern collapse, which has spawned more related studies. Among them, the study of (Gulrajani et al., 2017) has attracted attention by introducing the gradient penalty:

Gradient Penalty (GP): Is calculated by first interpolating between real and fake samples, and then computing the gradient of the critic's scores with respect to these interpolated samples. The penalty is the squared deviation of the gradient norm from 1, averaged across the batch. The final loss function is as follows:

$$\begin{aligned} L(\theta_C) =& \mathbb{E}_{x\sim P_r}[C(x)] - \mathbb{E}_{\tilde{x}\sim P_g}[C(\tilde{x})] + \\ & \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[(\|\nabla_{\hat{x}} C(\hat{x})\|_2 - 1)^2\right] \end{aligned} \tag{3}$$

where $\hat{x}$ is sampled uniformly along straight lines between pairs of real and generated samples, and λ is a hyperparameter that controls the strength of the penalty, which is set as default value 10 to ensure Critic's gradient comply with the Lipschitz constraint.

This strategy encourages the generator to produce samples that are realistic enough to receive high scores from the discriminator, while the discriminator is penalized for having a gradient norm far from 1, ensuring that it behaves like a smooth function (Gulrajani et al., 2017) that provides useful gradients to the generator throughout the training process.

2.2. Indicator Formation

As previously mentioned, following the training of the GAN, the generator can now represent the complete and continuous distribution of healthy samples found within the training set. Subsequently, the difference between the test signal and the healthy distribution need to be measured to quantify the degree of anomaly in a new signal.

However, this learned distribution is implicit, which means that it is impossible to explicitly write out the mathematical form of this learned data distribution. In the final implementation of AnoGAN (Schlegl, Seeböck, Waldstein, Schmidt-Erfurth, & Langs, 2017), this process is simplified to whether a similar signal can be sampled from the distribution of the

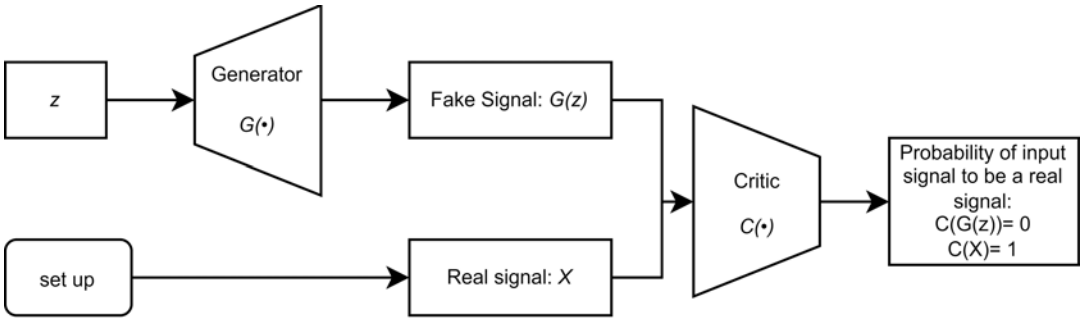


Figure 1. WGAN-GP Training Strategy

generated signal (D_{gen}), that is, finding z in the noise space (the distribution of Z, D_Z, is normally defined as a Gaussian distribution). This process is further reduced iteratively using back-propagation for a substantial number of iterations, such as 10,000 iterations, after which the final sampling result is considered the closest match to the test sample in D_{gen}.

However, the drawbacks of this process are evident; iterating multiple times for a single sample is computationally inefficient, especially when considering practical downstream applications. Moreover, using gradient descent optimization in isolation carries a significant risk of the sampling being trapped in local minima, which can adversely affect the quality of signal sampling in D_{gen}.

To enhance efficiency, the f-AnoGAN introduces an independent encoder for the sampling process. In GAN models, reliable mapping from D_z to D_{gen} is established. The aim of the independent encoder in f-AnoGAN is to facilitate the reverse process: mapping from the complex data distribution D_{gen} back to the simple feature space D_z. This process aids in quickly searching feature vectors z that match the new test sample best, enhancing both accuracy and efficiency in anomaly detection tasks. The obtained vector z is used to generate the corresponding health data $x_{gen} = G(z)$. The generated signal(x_{gen}) is then considered as the generated health signal closest to the tested signal to complete the corresponding indicator calculation.

To train the model, the encoder takes the generated signal X_{gen} as input and Z as output to train the parameters. The formation of the encoder can be depicted in Figure 2.

The loss of the training process of the Encoder is defined as:

$$\begin{aligned} \text{Loss} &= \text{Loss}_{\text{szs}} + \text{Loss}_f \\ &= \text{MSE}(X_{\text{gen}} - G(E(X_{\text{gen}}))) \\ &\quad + \text{MSE}(C(X_{\text{gen}}) - C(G(E(X_{\text{gen}})))) \end{aligned} \tag{4}$$

As mentioned earlier, the detection relies on the discrepancy between the test data and the generated healthy data. The discrepancies in this work are defined as two independent parts:

1. the Euclidean distance in the signal space
2. the Euclidean distance in the feature space, defined by the Critic

The Health Indicator (HI): the Anomaly Score (AS) is defined as the weighted sum of these two distances. In this research, this weighting parameter is not discussed emphatically and both distances are considered equally important, thus, for a tested data x, AS can be expressed as follows:

$$AS = \|x - G(E(x))\| + \|C(x) - C(G(E(x)))\| \tag{5}$$

2.3. Detection Part - Thresholding

The described method evaluates any signal to obtain a unique quantified indicator AS. For anomaly detection tasks, it is necessary to set a threshold based on the AS collection of the given healthy samples. The aforementioned method is applied to evaluate the healthy signals in the validation set to obtain the AS. Then, for the resulting Set_{AS}, the maximum likelihood estimation is used to estimate the parameters according to the assumed distribution type. In this step, research first establishes a distribution bank that includes all common and interesting distributions. Afterward, for Set_{AS}, different distribution estimates can be obtained, Dis_1, Dis_2, etc., along with the estimated distribution parameters. The AIC is taken as the matrix to compare and evaluate different distributions, and to select the optimal distribution based on this comparison as the parametric expression of the AS collection. The resulting dis is the distribution expression of the healthy signals. This distribution is inferred based on actual vibration measurements, and given the potential instability of operating conditions in the actual experimental process, and various interferences in signal measurements (such as the electrical environment), there are outliers in both the AS collection and the estimated distribution. This is also why similar studies do not use the maximum value of the validation's AS as the threshold for judging anomalies. In this method, the threshold setting is based on the estimation of the threshold according to the Distribution's Cumulative distribution function (CDF). In this paper, the AS corresponding to $CDF(\text{AS}) = 0.99$ is

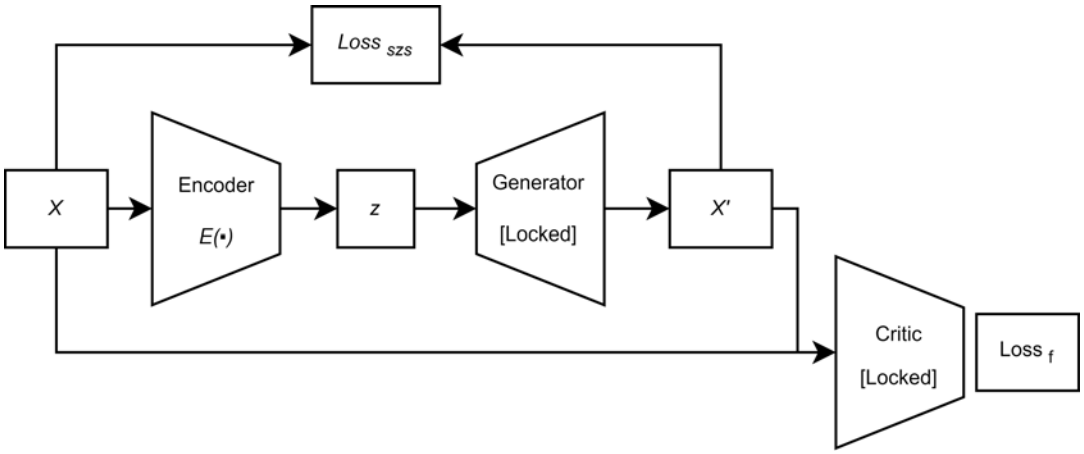


Figure 2. Encoder formation in f-AnoGAN

used as the threshold for judgment of anomaly.

3. Application of the Methodology and Results

3.1. Description of the data

The data used for validating the anomaly detection approach in this research were derived from a comprehensive gearbox degradation test (Van Maele et al., 2023).The measurements were conducted on an FZG multi-stage gearbox test rig (Figure 3), where the input and output of two gearboxes were mechanically interconnected, thus forming a mechanical closed loop. The load was provided by a friction disk coupling mechanism situated between the gearboxes, which applies torque to the gear meshing system through angular displacement between two discs at either end. Throughout the test, the torque, applied manually, was maintained at 60-90Nm, and the gear under test was set to a speed of 2560 rpm.

Figure 3. Photo of the multistage FZG test rig

Within the gearbox under investigation, the transmission system consists of three pairs of meshing gears, with their specific locations indicated in Figure 4. The test employed two pairs of helical gears made of 20MnCr5 steel, featuring 41 (monitored) and 38 teeth, respectively. Unlike standard, industrial gears, the gears under observation were not hardened (250HV) to ensure pitting would occur on the monitored surfaces within a reasonable time frame (Van Maele et al., 2023). A camera was used during the operation to periodically record the visual information of the gear surfaces at fixed intervals for the study and quantification of surface pitting. Specifically, the system was slowed down to 1rpm for image capture every 30 minutes during operation. The camera system took five shots of each meshing surface during the collection process, and the sharpest image was algorithmically selected to serve as the basis for quantifying the pitting area. Thus, after the test concluded, a quantitative description of the process of surface pitting area evolution over time was obtained, providing a metric for the degradation process.

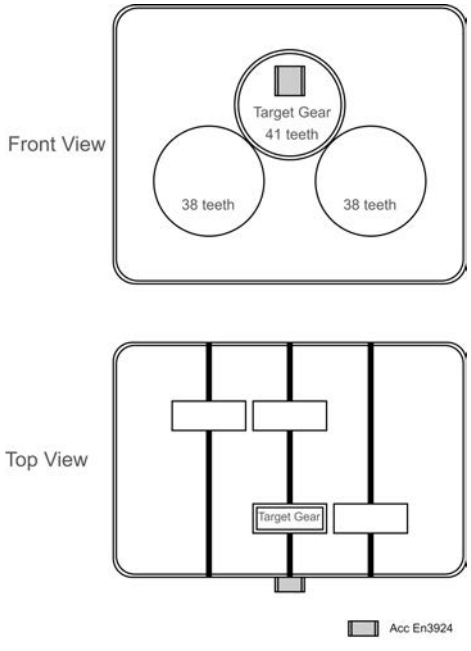


Figure 4. Sketch of the test gearbox setup

In addition to the visual information, during the test also torque,

speed, and most importantly, vibration signals were measured. The locations of the vibration sensors are indicated in the accompanying diagram. Vibration signals were sampled at 10-minute intervals, with each sample collected at a sampling rate of 25600Hz over a duration of 10 seconds. In total, this accelerated degradation test lasted approximately 205 hours.

Overall, the test yielded 999 valid vibration data entries (indicating failure after 999 cycles) and 300 sets of synchronized gear surface information. According to tribologists' analysis, significant and observable pitting occurred on all gear surfaces after the 33 cycle.Figure 5 is a depict of the observed degradation process of pitting.

In this study, the vibration signal in the X-direction from the vibration sensor EN-3924, which lies closest to the tested gear, was utilized. A health indicator built on a generative model was used to track the degradation of gears, aiming for earlier detection of pitting formation, which would, in turn, guide maintenance activities.

3.2. Data preprocessing

In this research, the training data were prepared with the following preprocessing steps to use more informative samples. It is postulated that the degradation features are primarily concentrated in the gear mesh frequency (GMF), its harmonics, and the features and structures of the sidebands. In this test, with the input shaft rotating at 2560 rpm, the velocity of the intermediate and target gears was calculated as $v_2 = v_1 \times \text{gear_ratio} = 39.5$Hz, and the gear mesh frequency was GMF $=$ Teeth_num $\times\ v_2 = 1622$Hz. Table 1 contains the characteristic frequencies of the test rig and the test.

Table 1. Characteristic frequencies of the test gearbox

Speed (Input Shaft)	2560 rpm
Speed (Driver Gear)	42.7 Hz
Speed (Target Gear)	39.5 Hz
GMF (Target Gear)	1622.6 Hz
2*GMF (Target Gear)	3245.2 Hz
3*GMF (Target Gear)	4867.8 Hz

Accordingly, the following preprocessing was applied to the data: initially, vibration signals were passed through a filter targeting the 1500-5000Hz frequency band, which includes the harmonics from the 1st to the 3rd order of the gear mesh frequency, along with the related band components. Afterward, the Discrete Fourier Transform was applied to isolate the informative frequency band, and then the 1500-5000Hz range was extracted to form the training, validation and test set samples.

The complete experimental dataset consists of approximately 999 independent measurements covering the full lifecycle. For this study, the early-life gear signals are selected as training samples to ensure that the training set comprised entirely healthy data to support model building and parameter optimization. A portion of the healthy dataset was also reserved as a validation set due to the encoder architecture of f-AnoGAN. In f-AnoGAN, unlike the original AnoGAN structure that relies solely on random sampling and gradient descent for sampling in D_{gen}, the model treats the training data as input to build the latent feature z via the encoder. Hence, to establish a threshold for anomaly detection, the new, unseen healthy data is still required as a reference.

Figure 6 delineates the division of the dataset in the test. Due to the run-in and gear bedding-in phases, which led to an unstable operational state of the experimental system, the initial two signals were discarded. The complete training set is composed of 19 independent signals, each sliced into time series of 51200 points with a 50% overlap during segmentation. All models in the study, including the generator, discriminator, and encoder, were trained exclusively with the aforementioned samples as per the described method. Furthermore, following the aforementioned method, 9 independent measurements from the healthy system were retained as a validation set, with the data division and sample generation being identical to the training set. All remaining data, encompassing both healthy and anomalous readings, were used as the final test set. It is important to note that the exclusion of data, as well as the delineation of the training, validation, and test sets, was conducted in chronological order following the accelerated degradation life course. In other words, the training and validation sets represent the early service life of the gears, while the test set includes the entire progression from a healthy state through the onset and development of pitting. Figure 7 illustrates the data pre-processing process.

3.3. Results

To verify the methodology's effectiveness, this study sets up comparative experiments and discusses the performance of the proposed model and the Autoencoder (AE). For the f-AnoGAN architecture, all models are set to be based on fully connected networks. To ensure fairness in the comparative experiments, the main model's structure and the number of parameters are kept as consistent and comparable as possible. This implies that both the generator of the proposed approach and the decoder of the AE, which serves as the benchmark method, undertake analogous functions by upscaling low-dimensional variables in the feature space to the target dimension. Consequently, to guarantee comparability between the two models, their parameters and network structures are configured to be identical. Both G and AE are completed by fully connected neural networks, transforming dimensions from $1000 \rightarrow 2500 \rightarrow 5000 \rightarrow 7001$, finally obtaining the spectrum from 1500-5000Hz (with a resolution of 0.5Hz). Based on these two models, model construction in

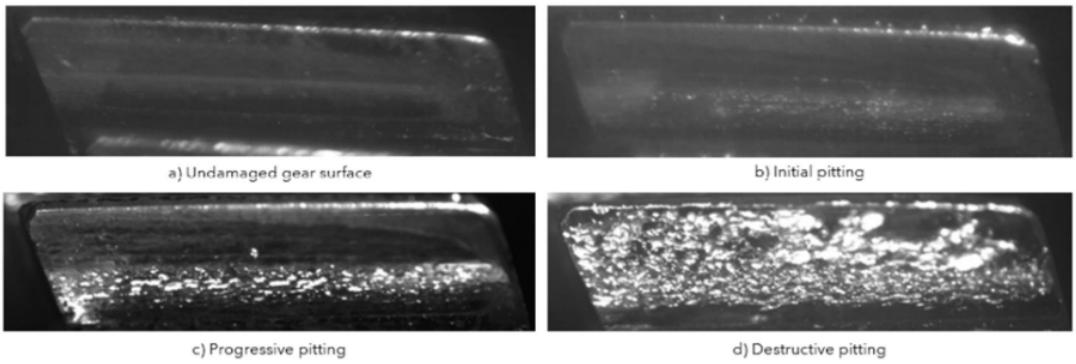


Figure 5. Observed degradation process of pitting

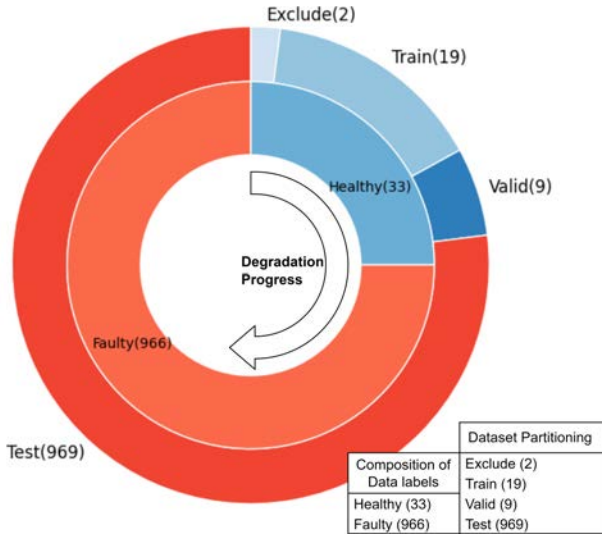


Figure 6. Segmentation of the training, validation, and test sets

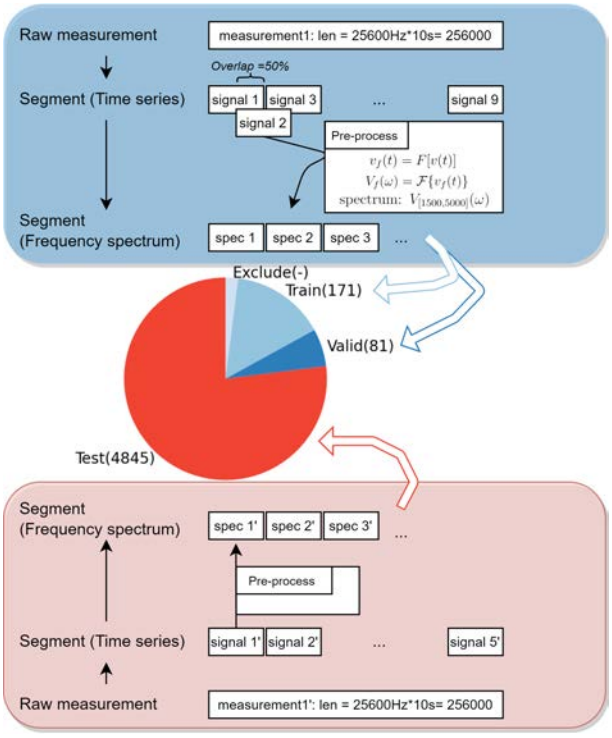


Figure 7. Dataset formation & pre-processing of signals, where $v(t)$ is the time-domain vibration signal, F is the filter, the Discrete Fourier Transform (DFT) of the filtered signal $v_f(t)$ is represented as $V_f(\omega)$.

each method is respectively completed, i.e., completing the Critic in GAN and implementing the encoder in f-AnoGAN. For the AE, the encoder is set according to the structure of the decoder. The AE uses the Mean Squared reconstruction Error (MSE) as its training loss and evaluates the health indicator during an assessment based on the MS reconstruction error.

For the proposed method, the training process generally follows the WGAN-GP (Wasserstein GAN with Gradient Penalty) scheme, setting the model to be trained for 2000 epochs (learning rate = 0.0001) to allow the model parameters to converge. After complete training, the generator can produce specified frequency bands based on any given set of features z. Figure 8 shows examples of training data and Figure 9 shows the generated results after 1990 epochs based on four randomly sampled z values. It is observed that the generated frequency bands closely mimic the features of the training data. From this, it can be inferred that the generator model has grasped the internal structure of the training data, that is, the representation of the service vibration condition of the given gear in the experimental system in the frequency domain. Specifically, the generated samples accurately replicate the gear mesh frequency and its higher-order harmonics, as well as the surrounding sideband performance.

According to the f-AnoGAN architecture, the construction of the Anomaly Score (AS) for any signal is completed with the help of the encoder. Figure 10 illustrates the result of generator sampling based on the aforementioned method and calculating the Euclidean distance in two spaces, with example

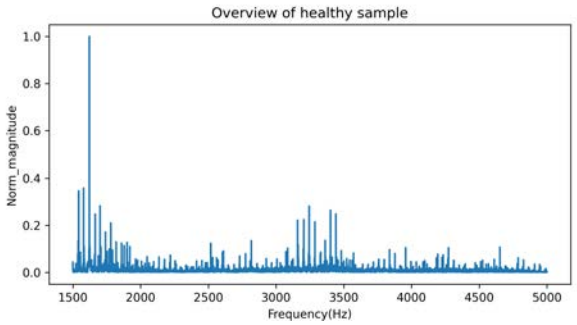


Figure 8. Sample overview in the training set

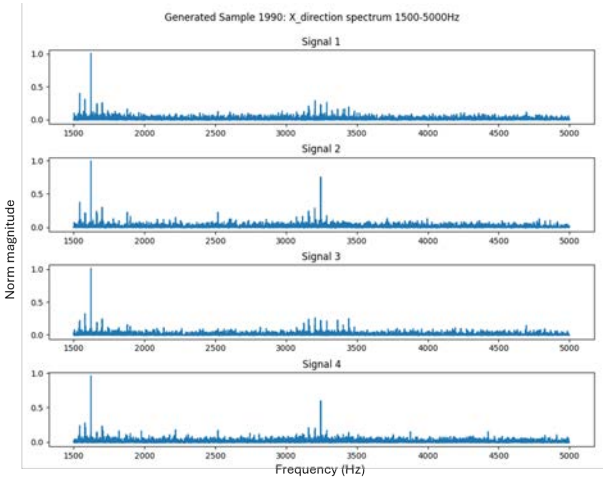


Figure 9. Generated spectrum at the end of the training period

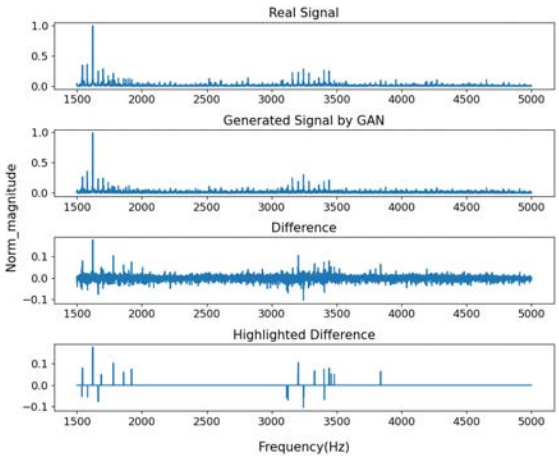


Figure 10. Comparison of the generated spectrum and the original spectrum

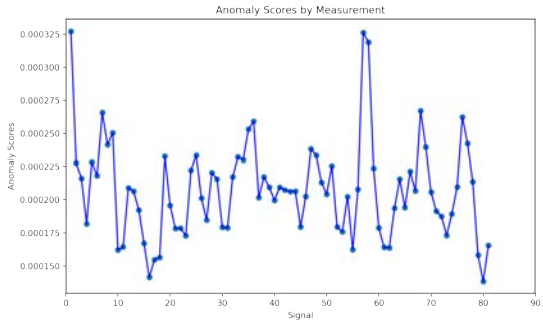


Figure 11. Anomaly score of the validation set (Healthy samples)

signals from the validation set.

Based on all the validation set data, a reference benchmark for anomaly detection, namely the Anomaly Score (AS) collection of healthy data, will be established. Figure 11 shows the AS for 81 samples (from 9 vibration signals) in the validation set 11. The study constructed a distribution bank using some common distributions. Following the aforementioned thresholding method, the fitting of the obtained distribution is as shown in Figure 12. The legend lists the distribution types in the figure, which are sorted in ascending order of the Akaike Information Criterion (AIC). Theoretically, a distribution with a smaller AIC value is closer to the true distribution. According to this criterion, the distribution among tested that best represents the AS of the healthy samples is the lognorm distribution (Figure 12). Additionally its numerical solution for various parameters is obtained, allowing to derive the CDF. Based on the aforementioned method where $CDF(AS_{\text{threshold}}) = 0.99$, the threshold is then determined for determining the anomaly (Figure 13). The threshold is then applied to the test set to evaluate the performance of the proposed method in anomaly detection.

Having completed all the components for anomaly detection as described previously, the Anomaly Score (AS) for each data in the test set is obtained following the aforementioned method. The average AS from the same vibration signal is taken as the AS for that vibration signal. Figure 14 shows the variation of AS across the entire test set in chronological order, along with the threshold performance. The results indicate that, according to the aforementioned method, the onset of failure occurs at the 33rd cycle (Figure 15), which aligns with the onset time of pitting derived from tribologists and visual information.

As a comparison, the AE model was also trained on the training set for 2000 epochs (learning rate = 0.0001), with an early-stopping at patience of 100. Figure 16 illustrates the reconstruction effect and schematic after completing the training.

As mentioned the reconstruction error derived from the AE model served as health indicators. In the comparative experiments, the threshold was established as the mean plus three times the standard deviation of the derived HI. AE-

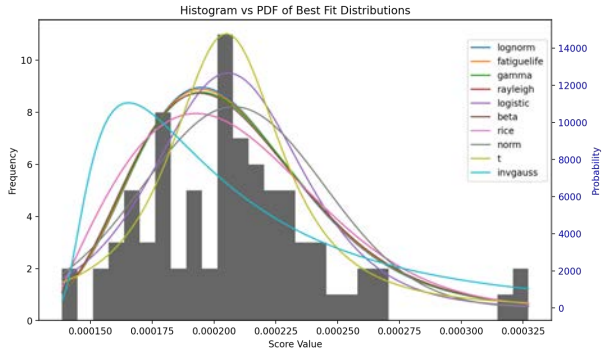


Figure 12. Distribution fit of the anomaly score in the Validation set

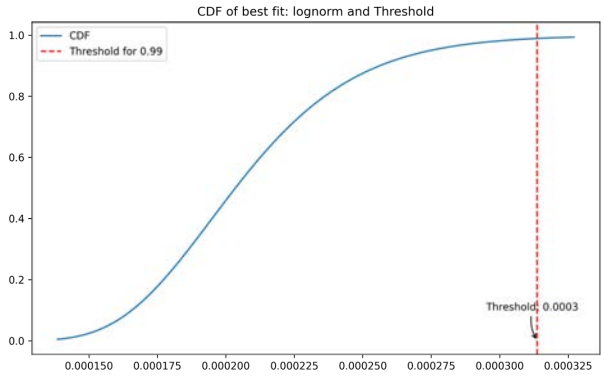


Figure 13. CDF according to the fitted distribution of the Anomaly Score of the healthy spectrum

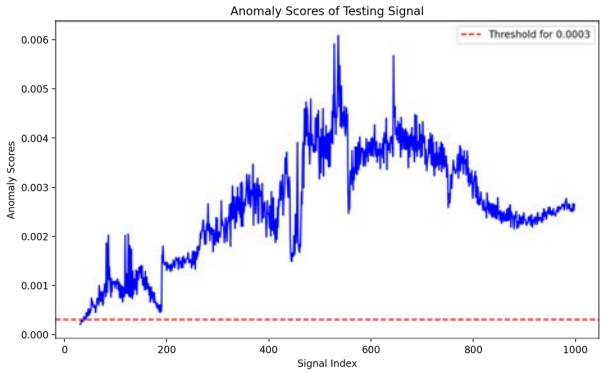


Figure 14. Detection result of f-AnoGAN-based anomaly detection method

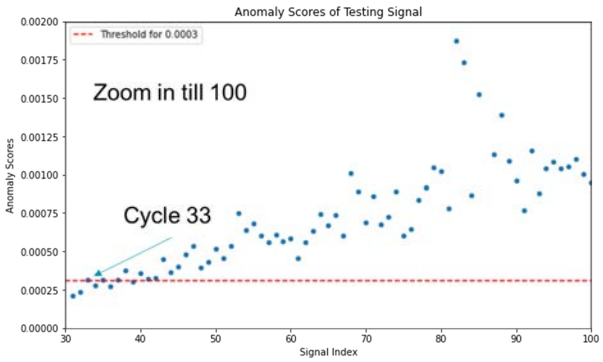


Figure 15. Anomaly Score in the first 100 cycles

based anomaly detection method identified the 38th cycle as the first cycle exceeding the threshold. However, this indicator was less stable, with a significant number of cycles between 38 and 200 falling below the threshold (Figure 18), indicating mis-detections if analogous to a classification problem.

3.4. Discussion

The results indicate that the proposed health indicator scheme and the thresholding method based on the GAN accurately detected the onset of the gear failure. Compared to the traditional unsupervised anomaly detection AE, the GAN-based detection advanced the detection time by 5 cycles. Given that the experiment conducted was an accelerated degradation test, and the gears underwent softening, this gap would be even more significant in actual industrial components.

Furthermore, it is also observed that the AE-reconstruction error, used as a HI, was highly unstable. One reason for this is the instability in the application of torque during the measurement campaign, which gradually diminishes during operation, necessitating the experiment to be halted and torque to be manually reapplied once excessive torque loss occurs. In the trend of HI obtained from AE, each sharp decrease in HI corresponds to the moments when the experiment is stopped and torque is reapplied. The same phenomenon is also observed in GAN-based HI. In GAN-based HI, even though HI is still established based on Euclidean distance, the powerful representation learning capability of the generator model allows it to construct more diverse samples that are more adaptable to certain instabilities in torque interference. Consequently, the differences reflected by HI are more attributable to degradation, with less impact from torque variations. This explains why GAN-based HI demonstrates better trend performance.

4. Conclusion

This work proposes an anomaly detection scheme for the condition monitoring of rotating machinery, focusing on gear fault detection using vibration signals. This method employs Generative Adversarial Networks (GANs) to learn the intrinsic structure and features of the training data's spectrum, particularly aiming to generate non-existent, highly realistic counterfeit samples. Based on the f-AnoGAN architecture, a health indicator is constructed utilizing the quantified Euclidean distance in two independent spaces. The study also employs a distribution fitting-based threshold method to assist in detection. The methodology is validated in a comprehensive

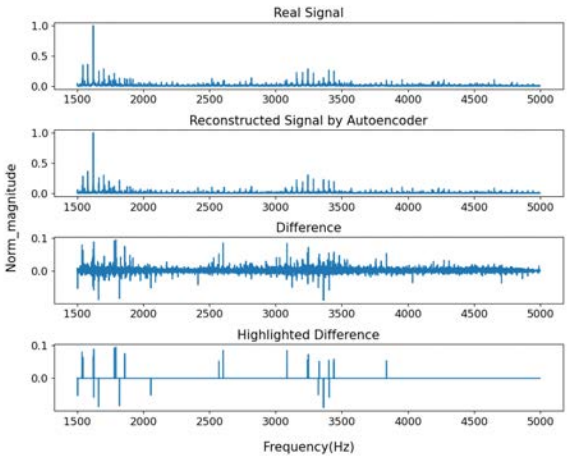


Figure 16. Comparison of the original spectrum and the reconstructed spectrum by an Autoencoder

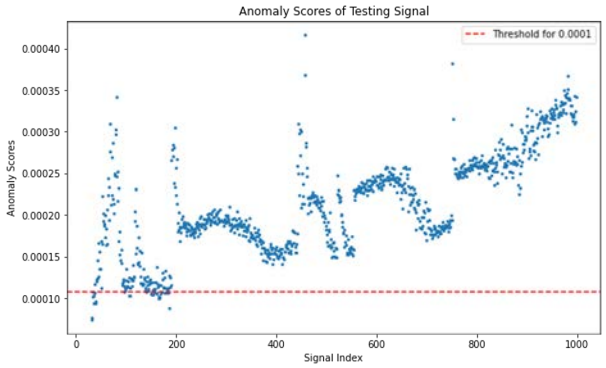


Figure 17. Detection result of Autoencoder-based anomaly detection method

gear accelerated degradation measurement campaign, which includes synchronized visual information collection, thus allowing for precise determination of the initial onset of pitting — the target of anomaly detection based on vibration signals in this research. In comparative experiments, the GAN-based method surpassed traditional unsupervised autoencoders and demonstrated better adaptability to changes in operating conditions, highlighting the performance of generative models with adversarial learning in the field of anomaly detection. Exploring how to better and more controllably utilize its adaptability under changing operating conditions will be the focus of future research.

Acknowledgment

This work was supported by Flanders Make, the strategic research center for the manufacturing industry, in the context of the QED project. Additionally K. Gryllias and Hao Wen would like to acknowledge the support of the FWO Fonds Wetenschappelijk Onderzoek – Vlaanderen in the frames of GOA3123N project. The computing resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

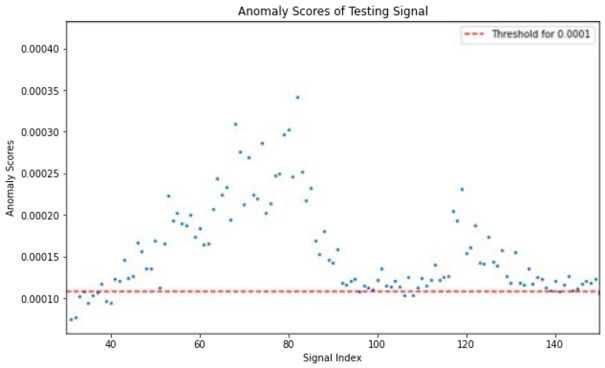


Figure 18. Autoencoder-based reconstruction error in first 200 cycles, first detection in cycle 38 and lots of misdetection in later cycles

References

Chen, H., Jiang, B., Ding, S. X., & Huang, B. (2022, March). Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 1700–1716. doi: 10.1109/TITS.2020.3029946

Dai, J., Wang, J., Huang, W., Shi, J., & Zhu, Z. (2020, October). Machinery Health Monitoring Based on Unsupervised Feature Learning via Generative Adversarial Networks. *IEEE/ASME Transactions on Mechatronics*, *25*(5), 2252–2263. doi: 10.1109/TMECH.2020.3012179

Ding, Y., Ma, L., Ma, J., Wang, C., & Lu, C. (2019). A Generative Adversarial Network-Based Intelligent Fault Diagnosis Method for Rotating Machinery Under Small Sample Size Conditions. *IEEE Access*, *7*, 149736–149749. doi: 10.1109/ACCESS.2019.2947194

Elasha, F., Ruiz-Cárcel, C., Mba, D., Kiat, G., Nze, I., & Yebra, G. (2014, July). Pitting detection in worm gearboxes with vibration analysis. *Engineering Failure Analysis*, *42*, 366–376. doi: 10.1016/j.engfailanal.2014.04.028

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017, December). *Improved training of wasserstein gans* (No. arXiv:1704.00028). arXiv. doi: 10.48550/arXiv.1704.00028

He, R., Tian, Z., & Zuo, M. J. (2022, April). A semi-supervised gan method for rul prediction using failure and suspension histories. *Mechanical Systems and Signal Processing*, *168*, 108657. doi: 10.1016/j.ymssp.2021.108657

Hendriks, J., Dumond, P., & Knox, D. (2022, April). Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, *169*, 108732. doi: 10.1016/j.ymssp.2021.108732

Kim, Y., Na, K., & Youn, B. D. (2022, March). A health-adaptive time-scale representation (htsr) embedded convolutional neural network for gearbox fault diagnostics. *Mechanical Systems and Signal Processing*, *167*, 108575. doi: 10.1016/j.ymssp.2021.108575

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014, January). Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mechanical Systems and Signal Processing*, *42*, 314–334. doi: 10.1016/j.ymssp.2013.06.004

Liu, C., & Gryllias, K. (2020, January). A semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis. *Mechanical Systems and Signal Processing*, *140*. doi: 10.1016/j.ymssp.2020.106682

Mao, W., Feng, W., Liu, Y., Zhang, D., & Liang, X. (2021, March). A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *150*, 107233. doi: 10.1016/j.ymssp.2020.107233

Öztürk, H., Sabuncu, M., & Yesilyurt, I. (2008, April). Early Detection of Pitting Damage in Gears using Mean Frequency of Scalogram. *Journal of Vibration and Control*, *14*(4), 469–484. doi: 10.1177/1077546307080026

Peng, D., Liu, C., Desmet, W., & Gryllias, K. (2023, July). Condition monitoring of wind turbines based on anomaly detection using deep support vector data description. *Journal of Engineering for Gas Turbines and Power*, *145*(091009). doi: 10.1115/1.4062768

Qin, Y., Wang, Z., & Xi, D. (2022, January). Tree cyclegan with maximum diversity loss for image augmentation and its application into gear pitting detection. *Applied Soft Computing*, *114*, 108130. doi: 10.1016/j.asoc.2021.108130

Ren, L., Sun, Y., Cui, J., & Zhang, L. (2018, July). Bearing remaining useful life prediction based on deep autoencoder and deep neural networks. *Journal of Manufacturing Systems*, *48*, 71–77. doi: 10.1016/j.jmsy.2018.04.008

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., … Kloft, M. (2018, July). Deep one-class classification. In *Proceedings of the 35th international conference on machine learning* (pp. 4393–4402). PMLR.

Salameh, J. P., Cauet, S., Etien, E., Sakout, A., & Rambault, L. (2018, October). Gearbox condition monitoring in wind turbines: A review. *Mechanical Systems and Signal Processing*, *111*, 251–264. doi: 10.1016/j.ymssp.2018.03.052

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In M. Niethammer et al. (Eds.), *Information processing in medical imaging* (pp. 146–157). Cham: Springer International Publishing.

Shi, J., Peng, D., Peng, Z., Zhang, Z., Goebel, K., & Wu, D. (2022, January). Planetary gearbox fault diagnosis using bidirectional-convolutional lstm networks. *Mechanical Systems and Signal Processing*, *162*, 107996. doi: 10.1016/j.ymssp.2021.107996

Teng, W., Wang, F., Zhang, K., Liu, Y., & Ding, X. (2014, January). Pitting Fault Detection of a Wind Turbine Gearbox Using Empirical Mode Decomposition. *Strojniški vestnik – Journal of Mechanical Engineering*, *60*(1), 12–20. doi: 10.5545/sv-jme.2013.1295

Van Maele, D., Poletto, J. C., Neis, P. D., Ferreira, N. F., Fauconnier, D., & De Baets, P. (2023). Online vision-assisted condition monitoring of gearboxes. In *8th european conference and exhibition on lubrication, maintenance and tribotechnology (lubmat 2023), proceedings.*

Wang, J., Li, S., Han, B., An, Z., Bao, H., & Ji, S. (2019). Generalization of Deep Neural Networks for Imbalanced Fault Classification of Machinery Using Generative Adversarial Networks. *IEEE Access*, *7*, 111168–111180. doi: 10.1109/ACCESS.2019.2924003

Xiang, L., Yang, X., Hu, A., Su, H., & Wang, P. (2022, January). Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. *Applied Energy*, *305*, 117925. doi: 10.1016/j.apenergy.2021.117925

Zhang, Y., Liu, W., Wang, X., & Gu, H. (2022, July). A novel wind turbine fault diagnosis method based on compressed sensing and dtl-cnn. *Renewable Energy*, *194*, 249–258. doi: 10.1016/j.renene.2022.05.085

Zhou, C., & Paffenroth, R. C. (2017, August). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 665–674). Halifax NS Canada: ACM. doi: 10.1145/3097983.3098052

Zhou, K., Diehl, E., & Tang, J. (2023, February). Deep convolutional generative adversarial network with semi-supervised learning enabled physics elucidation for extended gear fault diagnosis under data limitations. *Mechanical Systems and Signal Processing*, *185*, 109772. doi: 10.1016/j.ymssp.2022.109772

Zhu, R., Mousmoulis, G., & Gryllias, K. (2023, July). Wavelet-based high order spectrum for local damage diagnosis of gears under different operating conditions. In *Surveillance, Vibrations, Shock and Noise.* Toulouse, France: Institut Supérieur de l'Aéronautique et de l'Espace [ISAE-SUPAERO].

BIOGRAPHIES

Hao Wen received his Bachelor of Science degree from the China University of Geosciences, China and his Master of Science degree in Mechanical Engineering from the University of Manchester, UK. He joined the Noise and Vibration Research Group in the Department of Mechanical Engineering at KU Leuven, Belgium, as a PhD researcher in 2022. His research interests lie in the areas of fault detection utilizing unsupervised learning and generative modeling.

Djordy Van Maele Received his M. Sc. in electromechanical engineering technologies from Ghent University, Gent, Belgium, in 2021. In 2021 he started working on his doctoral degree in electromechanical engineering, in the field of tribology, at Ghent University until current date.

Jean Carlos Poletto obtained his M.Sc. degree in Mechanical Engineering from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2018. He has been developing research on experimental tribology, with active contributions to the field since 2015. Currently, he is working on a joint PhD program between UFRGS, Brazil, and Ghent University, Belgium.

Patrick De Baets received a master's degree in Electromechanical Engineering (1989) from Ghent University. His research, addressing the use of Thin Layer Activation to assess the fretting wear of machine components, resulted in a Ph.D. degree in 1995. He is now a full professor of machine elements and tribology. His research focusses on the tribological response (friction, friction stability and wear) in dry or marginally lubricated conditions of advanced materials such as fibre reinforced composites, high temperature ceramics and various self-lubricating materials. For his research, he has constructed different experimental tribological set-ups, ranging from the N to MN force range. The results of his research have been published in about 300 peer reviewed journal contributions and conference papers and numerous technical reports. Since more than 20 years Patrick De Baets is teaching specialised courses on Machine Design, Machine Elements and Tribology to Bachelor and Master students in Mechanical Engineering. Besides that, he has gained 15 years' experience in teaching general course on mechanical engineering to non-technical audiences, such as e.g. commercial engineering students.

Konstantinos Gryllias received the Diploma and Ph.D. degrees in mechanical engineering from the National Technical University of Athens, Athens, Greece, in 2004 and 2010, respectively. He is currently an Associate Professor of vibro-acoustics of machines and transportation systems with the Department of Mechanical Engineering, KU Leuven, Leuven, Belgium. He also serves as the Manager of the University Core Lab Flanders Make@KU Leuven Motion Products, Belgium. His research interests include condition monitoring, signal processing, prognostics, and health management of mechanical and mechatronic systems.

A Hybrid – Machine Learning and Possibilistic – Methdology for Predicting Produced Power Using Wind Turbine SCADA Data

Maneesh Singh

Western Norway University of Applied Sciences, 5020 Bergen, Norway
maneesh.singh@hvl.no

ABSTRACT

During its operational lifetime, a wind turbine is continuously subjected to a number of aggressive environmental and operational conditions, resulting in degradation of its parts. If left unattended, these degraded components will negatively influence its performance and may lead to failure of the wind turbine. In order to mitigate the risk associated with the failure of components, a wind turbine is regularly inspected and maintained.

Currently, there are two commonly used approaches for making maintenance management (inspection and maintenance) plans. Traditional Approach utilises understanding of failure profile of the components for manually developing maintenance plan for the equipment. Condition-Based Approach utilises data collected by condition monitoring of equipment for developing dynamic maintenance plan. SCADA system offers a low-resolution condition-monitoring data that can be used for fault detection, fault diagnosis, fault quantification and fault prognosis and eventually for maintenance planning.

The monitoring data from SCADA system of a wind turbine is often afflicted with variability and uncertainty. The variability in data is the result of continuously changing environmental conditions and uncertainty arises due to imperfections in the recorded data. The uncertainty may be due to many reasons, including, inherent characteristic of sensing devices, drift in calibration with time, deterioration of sensing devices' sensitivity and response due to environmental attacks, etc.

For handling variability in monitoring data a number of parametric and non-parametric (statistical) predictive models for different aspects of a wind turbine's structure and operation have been proposed. Depending upon its type – aleatory or epistemic – an uncertainty can be handled in a number of ways. Since, the dynamic nature of wind turbine operation does not allow collection of multiple values under the same condition; hence, uncertainty is mostly epistemic in nature. Possibilistic Approach, based on Fuzzy Set Theory, is especially suitable for handling epistemic uncertainty that may arise due to imprecision or lack of statistical data.

Aim of the ongoing research is to develop a methodology for detecting sub-optimal operation of a wind turbine by comparing Measured Produced Power against Predicted Produced Power. Unfortunately, variability and uncertainty associated with the recorded data make accurate prediction of produced power challenging.

This paper presents methodologies for predicting produced power using SCADA data while simultaneously accounting for variability and uncertainty. The methodologies utilise either parametric (example, power curve) or machine learning (example, XGBoost) models for handling variability; and Possibilistic Approach for handling uncertainty.

1. INTRODUCTION

1.1. Background

The world has two conflicting needs, on one side is the need to generate and supply more energy to bring people out of poverty and improve their living standard; on the other side is the need to reduce reliance on fossil fuel so as to cut down on emissions that cause global warming. These conflicting needs have acted as a spur to find economical and clean alternative sources of energy. In recent years, wind power has become one of the major sources of alternative energy and its share is expected to continuously grow in the coming decade (Global Wind Energy Council, 2021).

Due to various financial, social ("not-in-my-backyard" syndrome), environmental (meteorological conditions) and geographical (topological features) reasons the wind turbines are often located in remote areas where they experience harsh environmental conditions. The inconsistent and aggressive environmental conditions, like, wind velocity, humidity, temperature, precipitation and icing, degrade the vulnerable components. If left unattended, these degraded components

will result in deterioration of performance and at times failure. To prevent that from happening, maintenance of wind turbines is needed throughout their lifetime. It is estimated that maintenance costs comprise of a significant proportion (10-25%) of the total annual operational cost (Nilsson & Bertling, 2007).

Currently, there are two commonly used approaches for making maintenance management plans (tasks and schedules):

(a) **Traditional Approach** – In which understanding of the failure profile (failure causes, failure mechanisms, failure modes, failure rates, etc.) of components is used to develop maintenance concept and maintenance plan for the equipment.

(b) **Condition-Based Approach** – In which data, collected using condition-monitoring equipment or Supervisory Control and Data Acquisition (SCADA) systems is analysed for fault detection, fault diagnosis, fault quantification and fault prognosis and maintenance planning.

The Traditional Approach analyses structural, environmental and operational attributes to develop corrective or preventive maintenance plans. The preventive maintenance plans are often time-based, for example, preventive maintenance activities of wind turbines are normally planned at 3 to 6-month intervals based upon the age and condition of the turbine (Nilsson & Bertling, 2007). Since these time-based inspection and maintenance plans are expensive to execute, there have been efforts to develop methodologies based on formalized risk analysis, e.g., Risk Based Inspection and Maintenance or Reliability Centered Maintenance. This involves understanding failure profile and carrying out risk analysis & risk evaluation for preparing maintenance plans that are more efficient and effective than time-based or incidence-based maintenance plans (Fischer, Besnard & Bertling, 2012).

The Condition-Based Approach improves upon the inspection and maintenance plan by using condition attributes to update the equipment's risk assessment by detecting faults. This is achieved by (a) intermittent or continuous monitoring using sensors; (b) data analytics; and (c) developing condition-based maintenance plans. This approach can be applied using either (Tavner, 2012):

1. **Condition Monitoring System (CMS)** – A high-resolution specialized system for detailed analysis of the condition of a machinery by monitoring parameters like, speed, displacement, vibration and oil particles, using sensitive sensors. While specialized Condition Monitoring Systems can give accurate and detailed analysis, they are also expensive to install and use.
2. **Supervisory Control and Data Acquisition (SCADA)** – A low-resolution, usually at 10-minute intervals, standard system in every large wind turbine that monitors parameters for characterising environment, electrical, operational or structural attributes. SCADA system uses this data for controlling the wind turbine's operation after analysing its operating conditions and status. This data can also be used for deducing the health (fault detection, diagnosis and quantification) of the wind turbine.
3. **Structural Health Monitoring (SHM)** – A low-resolution system for monitoring health of a structure, including tower and foundation.

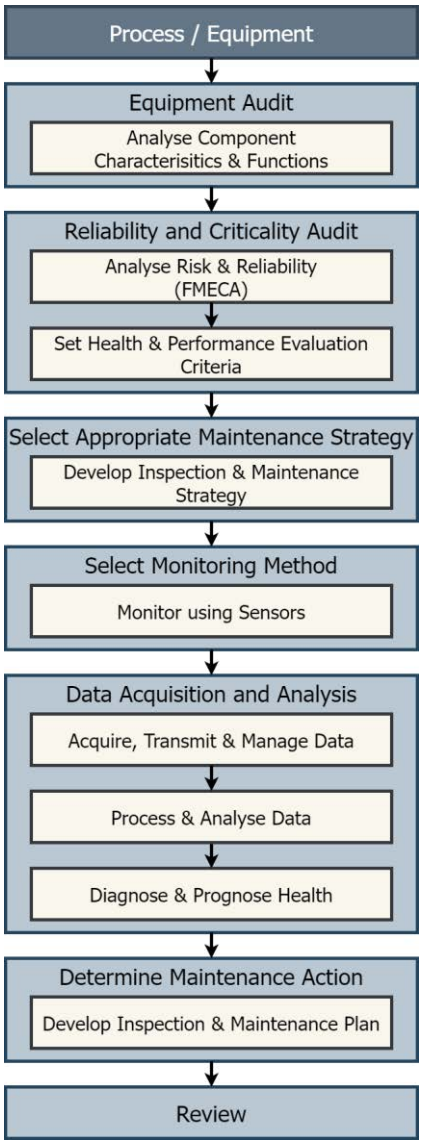


Figure 1. Main steps of a monitoring system (Based on ISO17359).

While Condition Monitoring System (CMS) provides costly but in-depth coverage, SCADA and Structural Health Monitoring (SHM) can provide cheap but wide coverage. Hence, a number of commercially available SCADA systems offer real-time data analysis, using statistical and artificial intelligence techniques, for fault detection of components. Yet, there is a need for better diagnostics, prognostics and control techniques using SCADA (Tavner, 2012; Yang et al., 2018).

Since, both – Traditional and Condition-based – Approaches have their own advantages and disadvantages, most of the maintenance planning is carried out by integrating the two approaches. The integration provides a solution that is robust, effective, and efficient. In an integrated method (Bindingsbø et al. 2023):

- failure analysis is carried out in the traditional manner, and then the results of failure profile is used judiciously to develop a maintenance strategy;
- time for inspection and maintenance of a component is adjusted based upon outcome of condition monitoring.

Figure 1 shows main steps that should be carried out to monitor a system according to ISO17359. According to the standard, condition-monitoring approach has three steps (Equipment Audit, Reliability and Criticality Audit and Select Appropriate Maintenance Strategy) that help in developing maintenance plan using the Tradition Approach. Thereafter, three more steps (Select Monitoring Method, Data Acquisition and Analysis and Determine Maintenance Action) help in improving the maintenance plan by incorporating knowledge of system's condition.

1.2. Supervisory Control and Data Acquisition (SCADA) System

An offshore wind turbine is subjected to severe variations in the environmental and operating conditions. To continuously monitor these variations all modern wind turbines come with a Supervisory Control and Data Acquisition (SCADA) system (Pandit & Wang, 2024).

In a SCADA system, a multitude of sensors constantly monitor various meteorological and operational parameters; and the data is transmitted, processed and stored in SCADA supervisory computers. The parameters that are monitored include (Manwell, McGowan & Rogers, 2009):

- **Position** – blade pitch angle, nacelle direction
- **Temperature** – nose cone, gearbox bearing, gearbox oil, hydraulic system oil, generator bearing, generator stator windings, generator split ring chamber, transformer, busbar section, inverter, controllers, VCP control boards
- **RPM** – rotor, generator
- **Hydraulic Characteristics** – pressure, reservoir level, flowrate
- **Environmental Characteristics** – wind speed, wind direction, temperature, humidity
- **Electrical Characteristics** – active power, reactive power, voltage, current, phase displacement, frequency

Apart from the data collected using sensors that are connected to a wind turbine, a number of data streams from nearby weather stations are also recorded.

The recorded SCADA data is analysed using different deterministic, probabilistic, Fuzzy Logic, Machine Learning, Artificial Neural Networks and Deep Learning approaches to detect, diagnose and quantify failures in the components. Information gained after analysis is used to control the process or operation (Manwell, McGowan & Rogers, 2009; Tavner, 2012).

Based on the data collected and analysed, a SCADA system can perform the following tasks (Manwell, McGowan & Rogers, 2009; Pandit & Wang, 2024):

1. **Controlling Operating Conditions** – SCADA uses the information regarding environment and grid to determine the appropriate operating conditions. It then controls the components (pitch angle, brakes, generator connection to the grid, etc.) so that the turbine operates according to the determined task schedule.
2. **Monitoring for Fault Detection** – SCADA uses the data from sensors (example, bearing temperature, hydraulic oil temperature, etc.) connected to critical components to monitor their behaviour and detect potential faults or spurious behaviour.
3. **Raising Alarm in Case of Faulty Behaviour** – If SCADA detects abnormal behaviour of a component it can raise alarm and notify the operator.
4. **Triggering Safety and Emergency Response** – In case of situations that can escalate into an accident, SCADA can disconnect turbine from the grid and activate brakes to isolate and shut down the operation.
5. **Integrating with Power Grid** – SCADA can control integration of individual wind turbine into the power grid, thereby contributing to feed and stabilisation.

1.3. Condition-based Maintenance Planning Using SCADA Data

The data acquired from SCADA can be used for fault detection, where a fault can be of various kinds, for example, degradation of components, failure of sensors, operation beyond safe operating limits, problems associated with grid. While it may be possible to detect some of these faults directly, for example, failure of sensors resulting in irrational readings, other faults may only be detected indirectly (Manwell, McGowan & Rogers, 2009).

Depending upon the type of fault, the time span between inception to potential failure could be between a few seconds (example, generator earth fault) to a few weeks (example, wear-out of gears). For the faults that have a long time span, analyses of SCADA data using appropriate models for fault diagnosis, fault quantification and finally fault prognosis may help in planning maintenance activities. These activities can be:

- triggered either when some condition indicator crosses a pre-set limit, or
- decided based on combination of Failure mode, Effect and Criticality Analysis (FMECA) with the condition analysis (fault diagnosis, quantification and prognosis) to update the existing maintenance plan.

The recommended maintenance activities may include inspection (visual, auditory, NDT), testing, service (lubrication, cleaning, repair, etc.), repair and replacement tasks. These activities may be either preventive or corrective in nature depending on whether the needed task is carried out before or after failure. Since maintenance activities are planned based on the actual monitored condition, condition-based maintenance strategy offers advantages that are associated with (Bindingsbø et al. 2023, Tavner, 2012):

- maintenance activities being carried out when required and not limited to corrective or preventive maintenance;
- not conducting unnecessary scheduled replacement of parts before their end of useful life.

In spite of these advantages, use of the Condition-Based Approach is still restricted and needs further research and development. This is because of the difficulties associated with the (Bindingsbø et al. 2023):

- quality and quantity of collected data,
- handling of imperfect (spurious, inconsistent, inaccurate, uncertain, or irrational) data collected from faulty sensors,
- interpretation of data for fault diagnosis, quantification and prognosis,
- updating of maintenance plan, and
- handling of unreliable analysis that may trigger false alarm (false positive) or failure to respond (false negative)

1.4. Methodologies for Predicting Produced Power

One of the common methods for analysing the performance of a wind turbine using SCADA data is to understand the power generation as a function of various variables, especially wind speed. A significant difference between the predicted power generation and measured power generation gives an indication of sub-optimal performance, hence, need for detailed examination. For this purpose it is essential to be able to accurately predict power generation under varying environmental and operating conditions (Pandit & Wang, 2024; Wang et al., 2016).

Power curve of a wind turbine is the unique relationship of a wind turbine between the power it generates and the environmental and operational conditions under which it operates. The power generated by a wind turbine is dependent upon the technical (example, radius of the rotor), environmental (example, wind speed, air density) and operational (example, pitch angle, angle between wind and nacelle) attributes (Manwell, McGowan & Rogers, 2009).

In a simplified power balance model, the wind power is converted to rotor power; which in turn is converted to electrical power. The efficiency of conversion of wind power to rotor power is dependent upon wind speed, air density, blade geometry, etc. Ideally, the rotor power should be converted entirely to the electrical power via its drive train system; but in reality, some power is lost as vibration and heat. The energy balance can be expressed as (Manwell, McGowan & Rogers, 2009):

$$P_{Rotor} = P_{Electrical} + P_{Vibration} + P_{Thermal} \quad (1a)$$

$$P_{Rotor} - P_{Electrical} = P_{Vibration} + P_{Thermal} \quad (1b)$$

Where:

P_{Rotor} = Rotor power
$P_{Electrical}$ = Electrical power
$P_{Vibration}$ = Vibration power
$P_{Thermal}$ = Thermal power

Hence, an increased discrepancy between rotor power (P_{Rotor}, predicted using models) and electrical power ($P_{Electrical}$, measured) is an indication of additional loss of energy due to increase in vibrations and heat generation-dissipation. This in turn can be attributed to the falling health condition of the mechanical and electrical drive train components. Thus, analysis of produced power can be used for (Duguid, 2018):

- **Fault Detection** – While exact cause may not be easy to identify, but a significant difference may help in fault detection necessitating further investigation.
- **Suboptimal Performance Detection** – Suboptimal performance, often due to poor control, can be identified using power curve. A comparison in power generation between a local group of wind turbines may also help in identifying those units that are performing sub-optimally.

To predict power generation, a number of parametric and non-parametric (statistical) methods have been proposed (Lydia at al. 2014; Pandit, Infield & Kolios, 2019; Saint-Drenan et al., 2020; Pandit & Wang, 2024). The parametric models are based on functions that correlate different variables and are of different types. For example, linearized segmented model, polynomial power curve, 4/5-parameter logistic function, etc. are based on power equation derived from Bentz's law, which can be expressed as (Manwell, McGowan & Rogers, 2009):

$$P_{Rotor} = P_{Wind} \times C_P(\lambda, \beta) \quad (2a)$$

$$P_{Electrical} = P_{Rotor} \times \eta \quad (2b)$$

$$P_{Electrical} = \left(\frac{1}{2}\rho A U^3\right) \times C_P(\lambda, \beta) \times \eta \quad (2c)$$

Where:

P_{Wind} = Wind power

η = Drive train efficiency ($generator\ power / rotor\ power$), (mechanical & electrical)

ρ = Air density

A = Rotor disc area

U = Air velocity

$C_P(\lambda, \beta)$ = Rotor power coefficient, it expresses the recoverable fraction of wind power and is a function of λ (tip speed ratio) and β (blade pitch angle).

The λ (tip speed ratio) can be expressed as:

$$\lambda = \frac{\Omega R}{U} \quad (3)$$

Where:

λ = Tip speed ratio

R = Radius of the wind rotor

Ω = Angular velocity (in radians/sec)

The maximum theoretically possible rotor power coefficient, $C_{P,max}$ also called the Betz limit, can be determined to be 0.59. The actual value of $C_P(\lambda, \beta)$ is much below the Bentz limit and is dependent upon technical features of the turbine and environmental factors (Saint-Drenan et al., 2020).

According to the **Equation 2c**, produced electric power is proportional to the density of air and cube of wind speed. The density of air is in-turn dependent upon the ambient temperature, humidity and pressure. It can be calculated according to:

$$\rho = \rho_d + \rho_v \quad (4a)$$

$$\rho_d = \frac{P - P_v}{\left(R_{Specific,Dry\ Air} \times T_k\right)} \quad (4b)$$

$$\rho_v = \frac{P_v}{\left(R_{Specific,Water\ Vapour} \times T_k\right)} \quad (4c)$$

$$P_{sat} = 6.1078 \times 10^{\frac{7.5T}{T+237.3}} \quad (4d)$$

$$P_v = \frac{(h \times P_{sat})}{100}$$

Where:

ρ_d = Density of the dry air

ρ_v = Density of the water vapour

T = Temperature (°C)

T_K = $T + 273.15$ (Kelvin)

h = Humidity

P = Total pressure of air

P_{sat} = Saturation water vapour pressure (Tetens' Formula)

P_v = Partial pressure of water vapour

$R_{Specific,Dry\ Air}$ = Specific gas constant for dry air = 287.05 J/(kg·K)

$R_{Specific,Water\ Vapour}$ = Specific gas constant for water vapour = 461.5 J/(kg·K)

The actual operation of a wind turbine is outcome of a number of controls, for example, aerodynamic torque control, yaw orientation control, brake torque control and generator torque control, that work together to create a number of decision combinations. The final operating strategy, which is an outcome of optimisation of diverse and often contradictory goals, determines the control of individual components. These goals include, safe operation, maximising power generation, minimising vibrations, preventing structural damages, integration with grid, etc. (Manwell, McGowan & Rogers, 2009).

Due to the complexities involved in accounting for all the parameters that can effect control and operation, the parametric models are often not accurate. Hence, for predicting power generation of existing wind turbines a number of models based on Artificial Intelligence (Support Vector Machine, Gaussian Process, Random Forest and Artificial Neural Network) have been propounded These models are trained using historical SCADA data and the trained models are later used for making predictions (Ouyang et al., 2017; Pandit, Infield & Kolios, 2019).

1.5. Data Quality for Predicting Power Produced

In spite of all the precautions, the measurements recorded by SCADA system are always afflicted with imperfections or uncertainties of various kinds. Where uncertainty of measurement can be defined as *the doubt that exists about the result of any measurement* (Bell, 1999).

Since, the uncertainties arise due to multiple reasons they are also of different types. Some of them are tangible (can be quantified), while others are intangible (cannot be properly quantified). Some uncertainties can by random and others can be systematic. Because of the difficulties associated with the taxonomy of uncertainties, a number of classifications have been proposed. Unfortunately, there is no consensus regarding these classifications and the proposed classifications have not been widely accepted, resulting in confusions. Traditionally, uncertainties have been classified into two types (Manwell, McGowan & Rogers, 2009; Simon, Weber & Sallak, 2018):

- **Aleatoric** – This type of uncertainty arises due to inherent randomness or variability of the measured parameter. By repeating the measurement, it is possible to express it in terms of mean and standard deviation (interval and confidence level).
- **Epistemic** – This type of uncertainty arises due to the lack of knowledge or data. The factors that contribute to the uncertainty influence all the recorded values, hence, there is limited benefit to be gained by repeated measurement. Epistemic uncertainty can be further classified into:
  - **Bias** – It is a systematic shift from the true value.

- **Inaccuracy** – This is the mean difference between the measured and true value of the measured variable.
- **Imprecision** – It refers to the length of interval between which the measured values lie.
- **Ignorance** – It arises due to limited availability of measurements or knowledge regarding precision.
- **Incompleteness** – It arises due to missing data.
- **Credibility** – It arises due to competence or trustworthiness during calibration, installation, etc.

Epistemic uncertainty can be evaluated based on information like the manufacturer's specifications, past experience, expert opinion or subjective feel.

For the sake of completeness, measurements should be reported along with their corresponding uncertainties. A tangible uncertainty can be quantified using two numbers: interval (width of margin of doubt or dispersion about the mean) and confidence level (confidence that the "true" value lies with that margin. Since the uncertainties of a measurement depends upon a number of factors, it is often difficult to quantify all of them (Bell, 1999).

These uncertainties are severe for wind turbines because of the large variations taking place in the environmental conditions. Most of the errors arise due to:

- **Imperfections Caused by Sensors** – These imperfections arise because of many reasons, including, variations in the parametric values, imperfect nature (bias, noise, etc.) of the instruments, incorrect calibration, drift in the instrument calibration, measurement location, etc. They may be characterised as:
  - **Inherent Imperfections** – Since, environmental conditions constantly change, the sensors report values based on their response time, sampling rate, resolution, sensitivity and statistical analysis. Each of these behaviour introduces different types of uncertainties.
  - **Acquired Imperfections** – During its operation, a sensor is exposed to a number of environmental attacks, like, variations in impacts, wind force, temperature, humidity, condensation, frosting / icing, vibrations, oil / dirt / salt deposition, etc., resulting in its degradation.
- **Imperfection Caused by SCADA System** – In a SCADA system, values are recorded every 10 minutes, hence, the recorded data is actually not of that particular time, but a statistical value based on predefined algorithm.

To ensure confidence in the data used for analysis, a number of corrective measures need to be taken. These include (Manwell, McGowan & Rogers, 2009; Tavner, 2012):

- **Use of High Quality Sensors** – High quality sensors should have structure that is able to withstand environmental attacks; and have superiority of performance in terms of accuracy, precision, reliability, repeatability and reproducibility.
- **Use of Multiple Data Streams** – Multiple and varied data streams can be used to confirm the same fault so that its probability of detection increases, for example, use of vibration and debris count for detecting bearing fault. Apart from the benefits of redundancy, use of different sensors at different locations increases the probability of detection. A negative side effect of this is the collection of excessive number of data streams resulting in data overload. Additionally, "law of diminishing return" dictates that use of multiple sensors for the same task may not provide any new information.
- **Use of Advanced Data Analytics Techniques** – A number of methods have been proposed to handle different types of uncertainties. While aleatoric uncertainty is often handled using the Probabilistic Approach, epistemic uncertainty can be handled using the Possibilistic Approach.

  In Possibilistic Approach, values are not regarded as "crisp point numbers" but as membership functions. By integrating Fuzzy arithmetic, that is based on extended interval analysis, with deterministic or Machine Learning models, the predicted output is not a crisp point but a Possibility Distribution Function. Comparison of this output membership function against acceptance criteria gives likelihood of failure in terms of "Possibility of Failure" and "Necessity of Failure". The advantage of using Possibility Distribution Function, over Probability Density Function, is that no preference is given to values within the range of Fuzzy interval. This suits well for the situations where the available data is sparse. The weakness of the Possibilistic Approach is its imprecise results, which may give over-conservative and, at times, uneconomical recommendations. Thus, Possibilistic Approach may be a useful tool for implementing the philosophy of zero tolerance of accidents where not only the probability but also any possibility of failure has to be eliminated (Ayyub & Klir, 2006; Ross, 2004).

2. MOTIVATION AND AIM OF THE RESEARCH

2.1. Motivation for the Research

As discussed in the previous section, performance of a wind turbine can be judged by comparing Predicted Produced Power and Measured Produced Power. A Significant Difference between the two indicates sub-optimal performance. **Figure 2** shows a flowchart of the methodology that can employed for detecting sub-optimal power production.

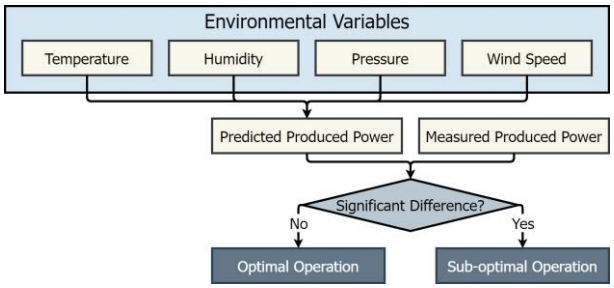


Figure 2. Flowchart showing the proposed fault detection methodology.

It may be possible to calculate Predicted Produced Power by using the four environmental variables; and if the Measured Produced Power (Grid Produced Power) is significantly less than the predicted value, there is a possibility that the wind turbine is operating sub-optimally.
While SCADA data can be used for carrying out this analysis, the methodology has some weaknesses. These weaknesses arise due to:

- lack of reliable models for calculating Predicted Produced Power taking into account all variations and imperfections in the collected data, and
- identification of what constitutes as *Significant Difference* considering the imperfections of the data.

2.2. Aim of the Research

Aim of the research is to develop a methodology for calculating Predicted Power Production using Hybrid (Machine Learning – Possibilistic) Approach while accounting for variability and uncertainty in the SCADA data.

2.3. Scientific Novelty and Importance of the Research

This paper presents work carried out to calculate Predicted Produced Power using wind turbine SCADA data using a Hybrid (Machine Learning – Possibilistic) Approach. The research includes:

- developing Machine Learning models for calculating Predicted Produced Power under varying environmental conditions, and
- handling of imperfections in the collected environmental and operating data by representing them as Fuzzy Membership Functions.

3. METHODS

3.1. SCADA Data Description

To demonstrate feasibility of the proposed methodology, SCADA data made available by the energy company EDP (2016) from four horizontal axis wind turbines located off the western coast of Africa has been used. The data has been recorded over a period of 2 years (2016 and 2017) at a 10-minute averaging interval. The datasets contain values of 76 parameters. For the mechanical components, some recorded parameters are (Bindingsbø et al. 2023):

- **Blades** – pitch angle
- **Rotor** – rpm
- **Nose Cone** – temperature
- **Nacelle** – direction, temperature
- **Generator** – rpm, bearing temperature (drive end and non-drive end), stator windings temperatures in the 3 phases, split ring chamber temperature, active power, reactive power
- **Gearbox** – bearing temperature, oil temperature
- **Hydraulic System** – oil temperature
- **High Voltage Transformer** – temperature
- **Ambient** – temperature, wind speed, wind direction

Associated dataset about meteorological conditions has also been provided for the same time instances. Failure logs containing timestamp, damaged component and associated remarks are also available. For this work, Turbine Number 7 ("T07") has been selected for which the total number of instances are 52445 and 52294 for 2016 and 2017, respectively. The variables that have been used in the calculation of power curve are given in **Table 1**.

Figure 3a shows the effect of Ambient Wind Speed on Generator RPM. The plot can be divided into three regions – (a) Low RPM Region, where Generator RPM < 300; (b) Transition Region, where 300 < Generator RPM < 1250; and (c) High RPM Region, where 1250 rpm < Generator RPM < 1680. When the Ambient Wind Speed is below the *Cut-In Wind Speed* (4 m/s), the frequency of Generator RPM below 300 rpm is high. With the increase in Ambient Wind Speed, the wind turbine adjusts its blade pitch angle so that Generator RPM is normally above 1250 rpm. Above the *Rated Wind Speed* (12 m/s), the Generator RPM is mostly above 1650 rpm. **Figure 3b** shows the effect of Ambient Wind Speed on Grid Produced Power. When the Ambient Wind Speed is below the *Cut-In Wind Speed* (4 m/s), Grid Produced Power is either negative or less than 275 kW. With increasing Ambient Wind Speed, Grid Produced Power increases so that at the *Rated Wind Speed* (12 m/s), Grid Produced Power is mostly *Rated Power* (2000 kW). **Figure 3c** shows the effect of Generator RPM on Grid Produced Power. The figure shows that the power generation drastically increases when the Generator RPM is above 1250 rpm.

Table 1. Selected variables used for developing the model.

Variable	Short Variable Name	Original SCADA Name	Description	Units
Timestamp			10-minute resolution	
Ambient Temperature	Amb_Temp	Amb_Temp_Avg	Average ambient temperature	ºC
Ambient Humidity	Amb_Humidity	Avg_Humidity	Average ambient relative humidity	%
Ambient Pressure	Amb_Pressure	Avg_Pressure	Average ambient pressure	millibar
Ambient Wind Speed	Amb_Wind_Speed	Amb_WindSpeed_Avg	Average windspeed within average timebase	m/s
Generator RPM	Gen_RPM	Gen_RPM_Avg	Average generator shaft / bearing rotational speed	rpm
Grid Produced Power	Grid_Prod_Power	Grd_Prod_Pwr_Avg	Power average	kW

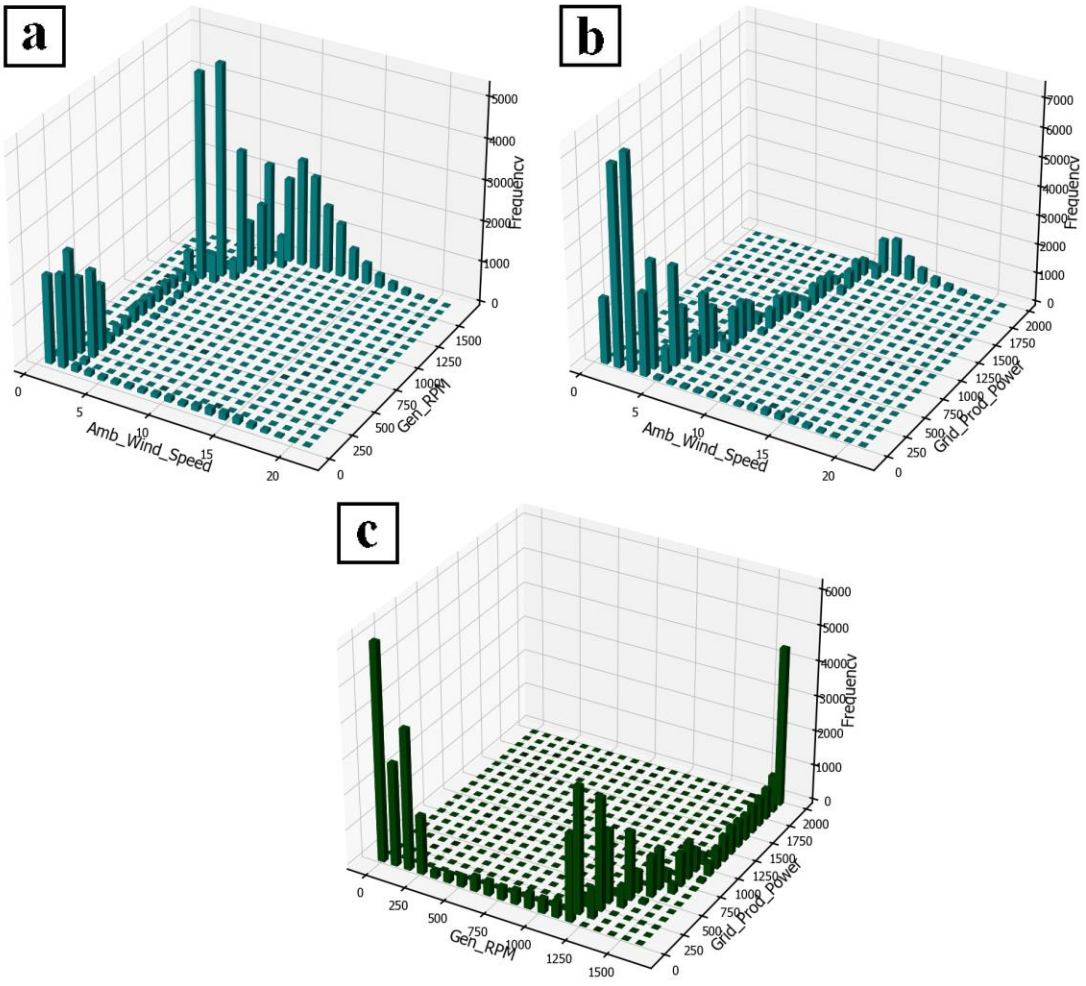


Figure 3. Relationships between Ambient Wind Speed, Generator RPM and Grid Produced Power.

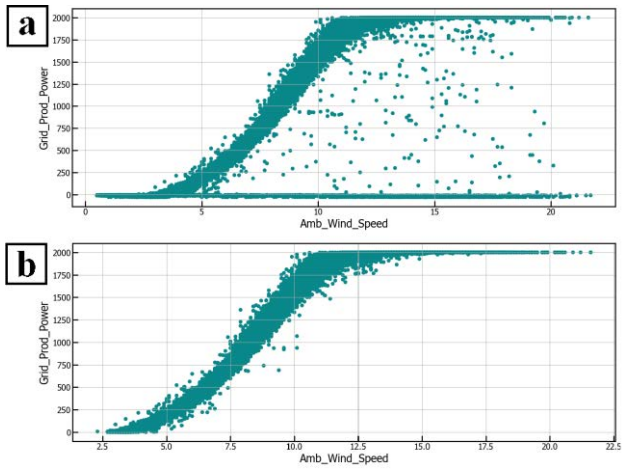


Figure 4. Plot of power generated versus wind speed using SCADA data. (a) Using raw data (b) Using data after removing outliers.

3.2. Data Pre-processing

Data pre-processing is an important step in the development of a Machine Learning model. This is to correct or remove vague, inconsistent, irrational, duplicate or missing values for algorithms to work properly (Bindingsbø et al. 2023).

SCADA data from a wind turbine also contain data that do not conform to the expected power curve and are referred to as "outliers". These outliers arise because of various explainable reasons. In this work, outliers have been identified for the following reasons:

Outlier Rule 1. *Generator RPM = 0 when Ambient Wind Speed => 4 m/s.* Even though the Wind Speed is above the *Cut-In Wind Speed* (4 m/s), the rotor does not move because the wind turbine is in the *shutdown state*. This can be because of various reasons, including the grid condition.

Outlier Rule 2. *Grid Produced Power <= 0 when Ambient Wind Speed < 4 and Generator RPM > 0.* This happens when the rpm of rotor is low, as a result of which power generation is less than the power consumed for operation. The difference is fulfilled by extracting power from grid.

Outlier Rule 3. *Grid Produced Power <= 0 when Ambient Wind Speed => 4 & Generator RPM > 0.* Even though the Wind Speed is above the *Cut-In Wind Speed* (4 m/s), the rotor is moving, power generation does not take place because the wind turbine is "free wheeling" in the *shutdown state*. This can be because of various reasons, including the grid condition.

Apart from these outlier data points, there are some more points that need to be removed. These data points have been recorded during the transition from normal operation to shutdown state or *vice versa*. These points lie scattered and can be identified using DBSCAN, a density-based clustering algorithm (Ester, Kriegel et al. 1996). Two rules that have been used for identifying the outliers are:

DBSCAN Clustering Rule 1. Ambient Wind Speed, Grid Produced Power, eps value = 2, min_samples value = 10

DBSCAN Clustering Rule 2. Ambient Wind Speed, Generator RPM, eps value = 3.45, min_samples value = 10

The results before and after cleaning are shown in **Figure 4**.

3.3. Flowchart for Predicting Produced Power

In order to develop a workable predictive model it is important to understand the process in terms of the structure, environment, and operation. **Section 1** briefly discusses some of these issues and based on this knowledge a simplified flowchart used for calculating Predicted Produced Power is shown in **Figure 5**. The figure also shows that there is a weak correlation between the environmental variables (Ambient temperature, Ambient Humidity and Ambient Pressure) and Grid Produced Power; but there is a strong correlation between Ambient Wind Speed and Grid Produced Power.

3.4. Representation of Variables as Possibility Distribution Functions

As discussed earlier, SCADA data is always encumbered by imperfections. One of the techniques that can be used for handling imperfections of the data is the Fuzzy Logic Approach. In this approach, a fuzzy variable X can be described by its Fuzzy Membership Function, instead of a Probability Density Function

In the Possibilistic Approach, a Fuzzy Membership Function can also be interpreted as a Possibility Distribution Function **(Figure 6)**. α–cut of this Possibility Distribution Function, donated by X_α, is a fuzzy interval $[x, x']$ that contains the values whose likelihood is α. The value of α can be in the range [0,1]. At the base, when the value of α is 0, variable has the interval within which the expected value will "certainly" lie. As the value of α increases, the interval between which the values lie decreases, but the certainty that the values will lie within this interval also decreases.

The α–cut of a fuzzy set is given by (Ayyub & Klir, 2006):

$$X_\alpha = [x, x']_\alpha = \{x \in X | x \leq x \leq x'\} \quad (5)$$
$$\alpha \in [0,1]$$

Where:

x = Lowest real number value of the interval

x' = Highest real number value of the interval

The use of α–cut allows for the concepts of interval analysis to be used (Ayyub & Klir, 2006).

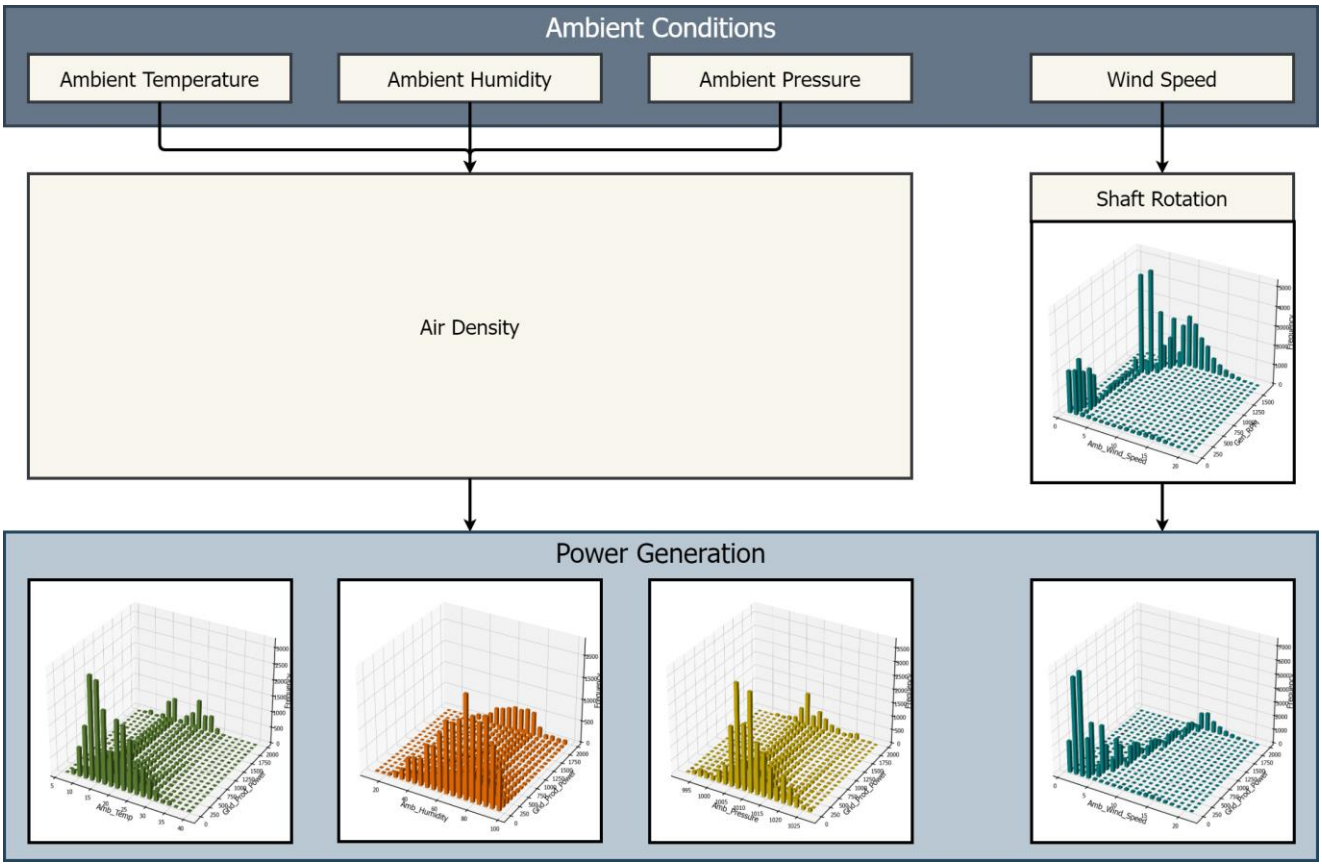


Figure 5. Flowchart showing influence of variables on the calculation of produced power.

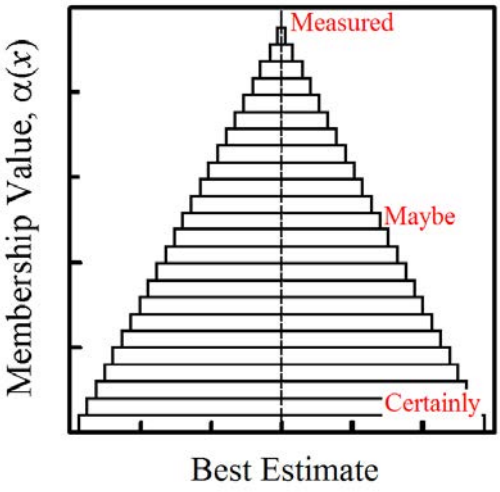


Figure 6. Conceptual illustration of possibility distribution function.

In the absence of detailed study to quantify the interval, limit values that have been used for the calculations are based on the literature and experience. For example, response time and uncertainty of a value recorded by a cup anemometer, depends upon its construction (dimensions, weight, etc.) and degree of deterioration (example, friction caused by corrosion). Under test conditions, a new anemometer can show inaccuracy of about 2%. Under working conditions, this inaccuracy may increase due to corrosion, wear, misalignment, deposition of dust, etc. (Manwell, McGowan & Rogers, 2009). Thus, at $\alpha = 0$ (interval within which the expected value "certainly" lies), the estimated limit of values around the measured values have been estimates as:

- Ambient Temperature : ±1.0°C
- Ambient Humidity : ±1.0%
- Ambient Pressure : ±1.0 millibars
- Ambient Wind Speed : ±0.5 m/s
- Power Coefficient : 0.45 ±0.05

Possibility Distribution Function for a variable is generated by stacking α number of intervals, where the bottom layer, $\alpha = 0$, has interval range:

Table 2. Possible combinations of interval values used for calculating Predicted Produced Power.

Combination	Ambient Wind Speed	Ambient Temperature	Ambient Pressure	Ambient Humidity
Combination_1	Min	Min	Min	Min
Combination_2	Min	Min	Min	Max
Combination_3	Min	Min	Max	Min
Combination_4	Min	Min	Max	Max
Combination_5	Min	Max	Min	Min
Combination_6	Min	Max	Min	Max
Combination_7	Min	Max	Max	Min
Combination_8	Min	Max	Max	Max
Combination_9	Max	Min	Min	Min
Combination_10	Max	Min	Min	Max
Combination_11	Max	Min	Max	Min
Combination_12	Max	Min	Max	Max
Combination_13	Max	Max	Min	Min
Combination_14	Max	Max	Min	Max
Combination_15	Max	Max	Max	Min
Combination_16	Max	Max	Max	Max

$$\left[\begin{matrix}(measured\ value - estimated\ limit\ value),\\ (measured\ value + estimated\ limit\ value)\end{matrix}\right]$$

In the Possibilistic Approach, in order to account for the uncertainty, instead of using crisp values of environmental variables (Ambient Temperature, Humidity, Pressure and Wind Speed) as recorded by SCADA and Power Coefficient, Possibility Distribution Functions of the variables are used. Calculations are carried out using interval values at each α–cut. For each value of α, the interval values of variables are determined. Considering all the minimum and maximum values of the intervals, the minimum and maximum values of the output function are calculated using accepted equations. Different combinations that are possible are shown in **Table 2**. The results of all α–$cuts$ are stacked to build the possibility distribution function of the output function (Ayyub & Klir, 2006).

3.5. Possibilistic Approach

The calculations are done in two steps. In the first step, Possibility Distribution Function for Air Density is generated using **Equation 4**. In the second step, the Possibility Distribution Functions for Air Density, Ambient Wind Speed and $C_P(\lambda, \beta)$ are used to generate Possibility Distribution Function for Predicted Produced Power using **Equation 2**.

3.6. Hybrid (Machine Learning – Possibilistic) Approach

Development of the Hybrid (Machine Learning – Possibilistic) is done in two steps.

In the first step, different Machine Learning models are trained using training dataset and the output from the trained models are evaluated. Models that have been evaluated are:

- Linear Models – Linear Regression (LR), Lasso, Ridge, and
- Tree-based Models – Decision Trees, Random Forest (RF)
- Boosting Models – AdaBoost, XGBoost and LGBoost
- Support Vector Regression (SVR)

Out of these models, XGBoost (RMSE = 186, R^2 = 0.93, MAE = 127) has been selected because it gives acceptable fit and takes short calculation time.

In the second step, the trained model and Possibility Distribution Functions of the environmental variables are used to generate Possibility Distribution Functions for Predicted Produced Power. The calculations are carried out according to the method described in the previous section, except that the calculations are done using the trained Machine Learning model instead of the equations.

4. RESULTS AND DISCUSSION

4.1. Possibilistic Approach

4.1.1. Effect of Environmental Variables on Air Density

Figure 7 shows the results of the calculations carried out for predicting Air Density. Since Air Density increases with the increase in Ambient Pressure, but decreases with the increase in Ambient Temperature and Ambient Humidity; the graph shows seasonal variations of the Air Density. The graph also shows sensitivity to the inaccuracies of recorded values and the "true" value may lie anywhere within the $Certain - Min$ and $Certain - Max$ curves.

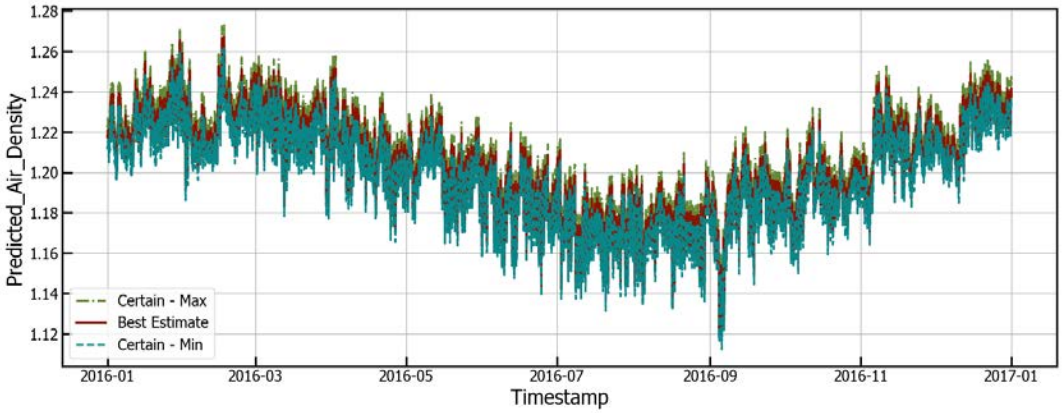


Figure 7. Seasonal variation on Predicted Air Density at $\alpha{-}cut = 0$.

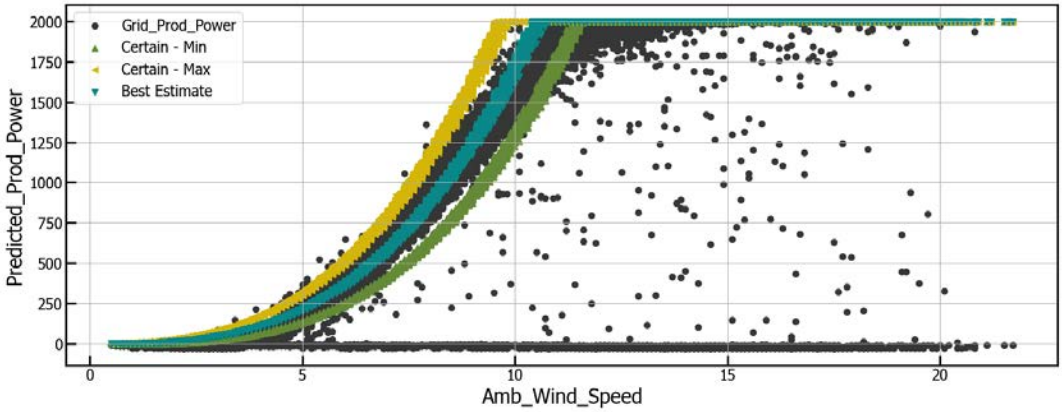


Figure 8. Effect of Ambient Wind Speed on Predicted Produced Power using Possibilistic Approach at $\alpha{-}cut = 0$.

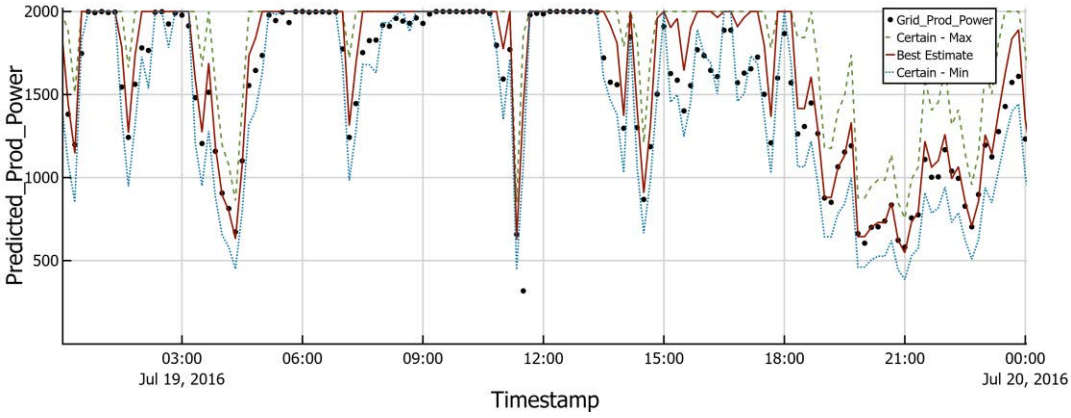


Figure 9. Plot of Grid Produced Power and Predicted Produced Power calculated using Possibilistic Approach at $\alpha{-}cut$=0 for a 24 hour duration (19th July, 2016).

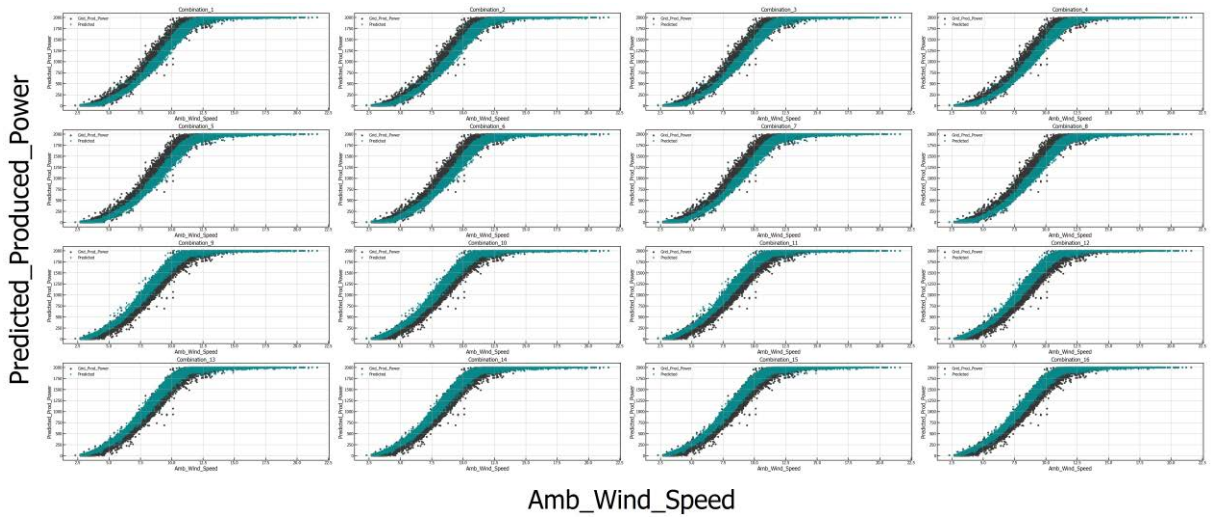


Figure 10. Predicted Produced Power using Hybrid Model for the combinations of interval values given in **Table 2** at α–cut=0.

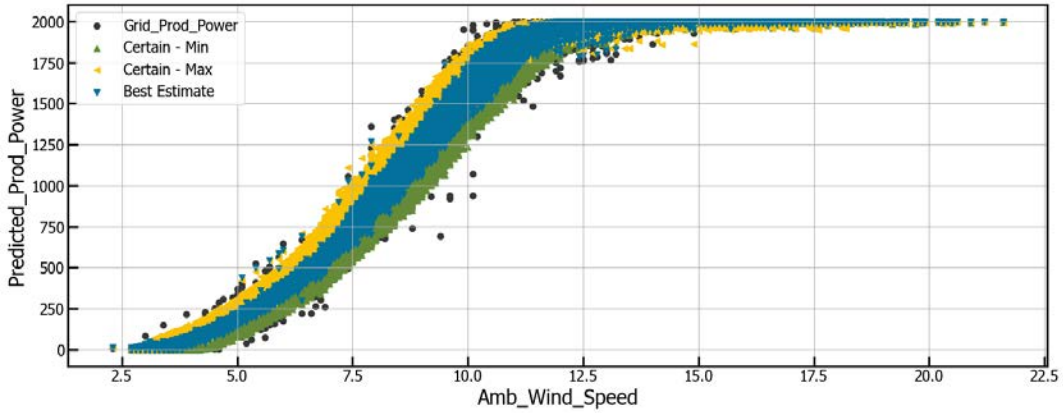


Figure 11. Effect of Ambient Wind Speed on Predicted Produced Power using Hybrid Approach at α–$cut = 0$. $Certain - Min$ is obtained from Combination_6 and $Certain - Max$ is obtained from Combination_11.

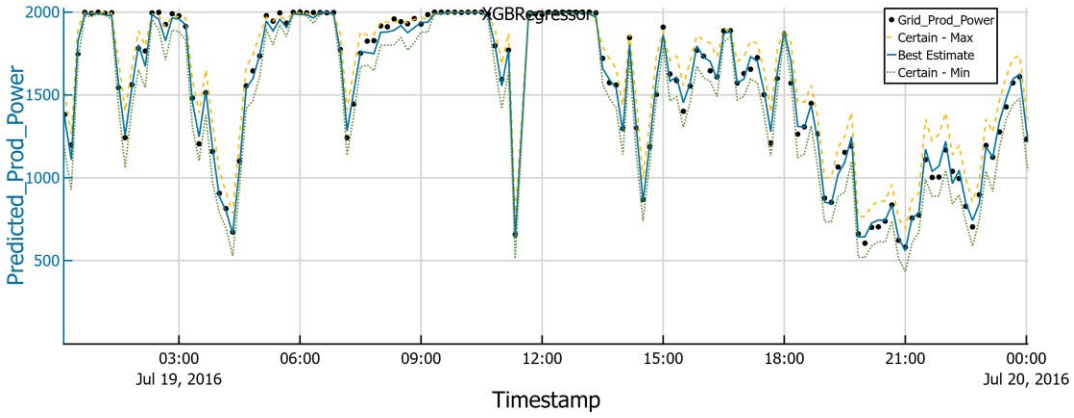


Figure 12. Plot of Grid Produced Power and Predicted Produced Power calculated using Hybrid Approach at α–cut=0 for a 24 hour duration (19th July, 2016).

4.1.2. Effect of Environmental Variables on Predicted Produced Power

Figure 8 shows the effect of Ambient Wind Speed on the Predicted Produced Power. The graph shows that:

- Power curve developed according to the **Equation 2** does not follow the actual trend. A better model, as proposed by Saint-Drenan, Y.-M. et al. (2020), may give better result.
- Spread of measured Grid Produced Power at a particular wind speed has not been accounted for. The spread can arise due to various reasons, like, control of the operation and imperfections in measurements.
- Predicted produced power is sensitive to the inaccuracies of recorded values and the "true" value may lie anywhere within the $Certain - Min$ and $Certain - Max$ curves.

Figure 9 shows plot of Predicted Produced Power and Grid Produced Power for a 24-hour duration (19th July, 2016). The graph shows that measured values generally lie within the boundaries set by $Certain - Min$ and $Certain - Max$ values.

4.2. Hybrid (Machine Learning – Possibilistic) Approach

Figures 10-12 show the results of calculations carried out using Hybrid (Machine Learning – Possibilistic) Approach. **Figure 10** shows the effect of max and min interval values of environmental variables on Predicted Produced Power. The figure shows that combinations have significant effect on the Predicted Produced Power.

According to **Equation 2**, Predicted Produced Power is proportional to cube of Ambient Wind Speed. Hence, Combination_1 to Combination_8 show lower values of Predicted Produced Power as compared to Combination_9 to Combination_16. Within these two sets of combinations, the differences are small because of the relatively small differences in the calculated air density.

Figure 11 shows the effect of Ambient Wind Speed on Predicted Produced Power using Hybrid Approach at $\alpha\text{–}cut = 0$. The figure shows significant effect of measurement uncertainties on the predicted values. $Certain - Min$ is obtained from Combination_6 and $Certain - Max$ is obtained from Combination_11.

Figure 12 shows plot of Grid Produced Power and Predicted Produced Power calculated using hybrid approach at $\alpha\text{–}cut$=0 for a 24-hour duration (19th July, 2016). The graph shows that measured values generally lie within the outer most boundaries set by $Certain - Min$ and $Certain - Max$ values.

A comparison between **Figure 9** and **Figure 12** shows that, in general, (a) Machine Learning model fits better than the parametric model; and (b) the difference between $Certain - Max$ and $Certain - Min$ in the Hybrid Model is less than that in the Possibilistic Model.

5. CONCLUSIONS

This paper presents a simple yet robust methodologies for calculating Predicted Produced Power using SCADA data while accounting for variability and uncertainty. The methodologies utilise either parametric or Machine Learning models for handling variability; and Possibilistic Approach for handling uncertainty. As a case study, the idea has been demonstrated using real-life SCADA data.

To take the research work further, the following tasks have been identified:

- The models do not account for effect of control measures of the wind turbine on produced power. Since, these measures can significantly effect power generation (López-Queija et al., 2022); models that account for control measures need to be used.
- Grid Produced Power has been assumed to have crisp values, but in reality measurement of Grid Produced Power is also afflicted with uncertainties. Hence, calculations need to be done by representing it by a Possibility Distribution Function.
- Having obtained Possibility Distribution Functions of Predicted Produced Power and Grid Produced Power, *Likelihood of Sub-optimal Performance* can be determined using the concepts of Possibility and Necessity Measures.

DATA AVAILABILITY

The datasets presented in this study can be found in online repositories given below:

- https://www.edp.com/en/wind-turbine-scada-signals-2016
- https://www.edp.com/en/innovation/open-data/wind-turbinescada-signals-2017.

REFERENCES

Ayyub, B.M. and Klir, G.J. (2006). Uncertainty Modeling and Analysis in Engineering and Sciences, Chapman & Hall/CRC Press, Boca Raton

Bell, S. (1999). A Beginner's Guide to Uncertainty of Measurement. Issue 2, National Physical Laboratory, Report No. 11

Bindingsbø, O.T., Singh, M., Øvsthus, K. and Keprate, A. (2023). Fault Detection of a Wind Turbine Generator Bearing Using Interpretable Machine Learning, Frontiers in Energy Research, 11:1284676, doi: 10.3389/fenrg.2023.1284676

Duguid, L. (2018), Data Analytics in the Offshore Wind Industry – Pilot Case Study Outcomes, CATAPULT - Offshore Renewable Energy Report No. PN000229-RPT-001. https://ore.catapult.org.uk/wp-content/uploads/2018/05/Data-Analytics-in-Offshore-Wind-Pilot-Case-Study-Outcomes.pdf

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996), A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, August 2-4, p. 226–231

Fischer, K.; Besnard, F.; Bertling, L. (2012). Reliability-Centered Maintenance for Wind Turbines Based on Statistical Analysis and Practical Experience, IEEE Transactions on Energy Conversion, Vol.27 (1), p.184-195

Global Wind Energy Council (2021). "Global Wind Report 2021", available at: https://gwec.net/wp-content/uploads/2021/03/GWEC-Global-Wind-Report-2021.pdf

López-Queija, J., Robles, E., Jugo, J., Alonso-Quesada, S. (2022). Review of Control Technologies for Floating Offshore Wind Turbines, Renewable and Sustainable Energy Reviews Vol. 167, 112787

Lydia, M., Kumar, Suresh Kumar, S., Selvakumar, A. I., Prem Kumar, G. E. (2014). A Comprehensive Review on Wind Turbine Power Curve Modeling Techniques, Renewable & Sustainable Energy Reviews, Vol. 30, pp.452-460

Manwell, J. F., McGowan, J.G. and Rogers, A.L. (2009). Wind Energy Explained — Theory, Design and Application (2nd ed.), John Wiley & Sons Ltd., ISBN 978-0-470-01500-1

Nilsson, J., and Bertling, L. (2007). Maintenance Management of Wind Power Systems Using Condition Monitoring Systems — Life Cycle Cost Analysis for Two Case Studies. IEEE Transactions on Energy Conversion, Vol. 22 (1), 223–229

Ouyang, T., Kusiak, A., He, Y. (2017). Modeling Wind-Turbine Power Curve: A Data Partitioning and Mining Approach, Renewable Energy, Vol. 102, pp. 1-8

Pandit, R. and Wang, J. (2024). A Comprehensive Review on Enhancing Wind Turbine Applications with Advanced SCADA Data Analytics and Practical Insights, IET Renewable Power Generation, Vol. 18, pp. 722-742

Pandit, R. K., Infield, D. and Kolios, A. (2019). Comparison of Advanced Non-Parametric Models for Wind Turbine Power Curves, IET Renewable Power Generation, Vol. 13(9), pp. 1503-1510

Ross, T. J. (2004). Fuzzy Logic with Engineering Applications, John Wiley and Sons Ltd, ISBN 9780470860748

Saint-Drenan, Y.-M. et al. (2020). A Parametric Model for Wind Turbine Power Curves Incorporating Environmental Conditions, Renewable Energy, Vol. 157, pp. 754-768

Simon, C., Weber, P. and Sallak, M. (2018). Data Uncertainty and Important Measures, John Wiley & Sons, EBOOK ISBN 9781119489351

Tavner, P. (2012). Offshore Wind Turbines — Reliability, Availability and Maintenance, The Institution of Engineering and Technology, IET Renewable Energy Series 13, ISBN 978-1-84919-230-9

Wang, S., Huang, Y., Li, L., Liu, C. (2016). Wind Turbines Abnormality Detection Through Analysis of Wind Farm Power Curves, Measurement, Vol. 93, pp. 178–188

Yang, W., Wei, K., Peng, Z. and Hu, W. (2018). Chapter 7, Advanced Health Condition Monitoring of Wind Turbines, W. Hu (ed.), Advanced Wind Turbine Technology, Springer International Publishing AG

A maturity framework for data driven maintenance

Chris Rijsdijk[1], M.J.R. van de Wijnckel[1,2], Tiedo Tinga[1,2]

[1]*Netherlands Defence Academy, Faculty of Military Sciences, P.O. Box 10000, 1780CA, Den Helder, The Netherlands*
c.rijsdijk.01@mindef.nl

[2]*University of Twente, Faculty of Engineering Technology, P.O. Box 217, 7500AE, Enschede, The Netherlands*
m.j.r.vandewijnckel@utwente.nl
t.tinga@utwente.nl

ABSTRACT

Maintenance decisions range from the simple detection of faults to ultimately predicting future failures and solving the problem. These traditionally human decisions are nowadays increasingly supported by data and the ultimate aim is to make them autonomous. This paper explores the challenges encountered in data driven maintenance, and proposes to consider four aspects in a maturity framework: data / decision maturity, the translation from the real world to data, the computability of decisions (using models) and the causality in the obtained relations. After a discussion of the theoretical concepts involved, the exploration continues by considering a practical fault detection and identification problem. Two approaches, i.e. experience based and model based, are compared and discussed in terms of the four aspects in the maturity framework. It is observed that both approaches yield the same decisions, but still differ in the assignment of causality. This confirms that a maturity assessment not only concerns the type of decision, but should also include the other proposed aspects.

1. INTRODUCTION

Von Leibnitz already dreamt of a universe where decision problems were solved by computations rather than by furious debates. Centuries later, it is much better understood that Von Leibnitz's dream cannot come true. So, one may compute many decisions, but not any decision. Where computed engineering decisions fail, maintenance decisions are typically triggered. Unsurprisingly, maintenance decisions are often hard to compute, or they may even be fundamentally incomputable. However, an inability to compute a decision does not imply that such a decision cannot be supported by computations. This paper will present a maturity framework for computational maintenance decision support.

In this framework, maturity grows as more (advanced) decisions in a maintenance control loop are computed. However, the presented framework not only considers the type of decision, as in existing data maturity models, but relates maturity also to: (*i*) the translation of reality to data (vice-versa), (*ii*) the computability (with models) of the decisions involved and (*iii*) the causality of the relations obtained. A case study will be used to explore the attainable maturity starting from the lowest level. An experience based and a model based approach will be attempted, which both will prove to take the correct decision for an arbitrary validation set. Still, decision makers should care about the approach as causality is managed differently. In the experience based approach, causality will be assigned afterwards. In the model based approach, causality is inherent, as a model that is posited as true is solved. Further, it is observed that it is impossible to compute a true model from only a history of measurements. Therefore, a history of measurements will be indecisive about the approach. Still, the engineering profession established a plethora of guidelines that have often proved to be correct. As these engineering guidelines strengthen (a suspicion of) causality for both approaches, the attainable maturity in data driven maintenance may rise at an acceptable risk.

This paper is organized as follows. Section 2 will introduce the four basic aspects of the framework to assess the maturity in data driven maintenance. Section 3 will portray a typical construction of two different autonomous fault detection and isolation methods (the first step in maturity). Section 4 will demonstrate fault detection and isolation in an iconic case study. Finally, section 5 will discuss the results and section 6 will present the conclusion.

2. BACKGROUND

This section will introduce the four basic elements that jointly determine the maturity in data driven maintenance and thus

constitute the proposed framework. Section 2.1 addresses the challenges in computing a "real" decision, section 2.2 will discuss the challenges in using (engineering) models to compute decisions. Then section 2.3 will relate the flow of the maintenance control loop with a conventional data maturity model. Finally, section 2.4 discusses the difference between observed associations and causality, and its effect on decision making.

2.1. Obstructions in computing "real" decisions

Data (Latin: givens) are input symbols to a syntactical formal language. Hilbert dreamt of a formal language that could provide a complete, consistent, and decidable foundation of mathematics. Gödel (1931), Church (1936) and Turing (1937) showed that such a formal language is nonexistent and the dreams of Von Leibnitz and Hilbert were destroyed. This means that some problems are fundamentally incomputable. Moreover, even the most potent computing devices may just fail to compute a problem in time. Therefore, problems that are computable in principle may be too complex to compute in practice.

A formal language becomes meaningful by assigning a truth value. Then, a computation may become similarly meaningful and it may eventually represent some reasoning about truth or falsehood. Evidently, Von Leibnitz similarly hoped to compute meaningful decisions as he hoped to settle legal disputes this way. Then, a computed decision involves

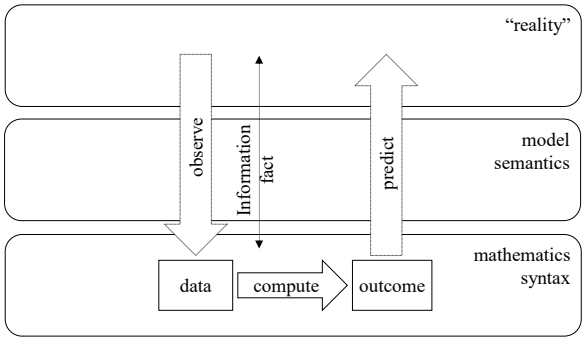


Figure 1. Framework for computing "real" decisions.

Engineers are not necessarily orthodox positivists but the engineering profession is not just about modelling, it also includes building. To circumvent or at least alleviate philosophical controversy, the "reality" in Figure 1 could also be seen as merely a user interface by skeptics who doubt the existence of space-time (Hoffman, 2019).

Any vertical translation in Figure 1 involves an arbitrary human choice, i.e. facts are *made* (Latin: facere) and information is *shaped* (Latin: formare). So, facts and information do not follow from some computable coding operation, they involve arbitrary human choice. For example, observing is not just a mechanical decoding of sound or light waves, it also involves a specific interpretation. Likewise, predicting involves more than just computing an outcome (100101…).

In conclusion, computing a "real" decision may be impossible because (*i*) it is fundamentally incomputable, (*ii*) it is too complex to compute in time, or (*iii*) the translation between "reality" and the syntactical computation is philosophically controversial.

2.2. Maintenance decisions are incomputable

A decision (Latin: cut-off) is the elimination of outcomes that would have occurred otherwise. A computation is a deterministic discrete operation that can be performed on a Turing Machine. In a way, a Turing Machine decides as it halts at a particular outcome (while eliminating all other candidate outcomes). So, syntactical decisions include the acceptance or rejection of a string as a well formed formula in a formal language. However, "real" decisions include a choice that causes a specific outcome, rather than any other outcome.

The computation of a "real" decision requires translations between a syntactical Turing Machine and "reality" (Figure 1). These translations are essential for data driven maintenance where computations from syntactical data should support "real" maintenance decisions. Generally, the engineering profession established a high degree of common sense regarding these potentially controversial translations. This common sense has been made explicit in guidelines that specify the computation of the quality of a design (CEN, 2007), (IACS, 2024). Quality is defined by ISO (2015):

> The degree to which a set of inherent characteristics of an object fulfils requirements.

So, quality reflects a margin between measurable inherent characteristics and subjective requirements. So, quality is not just a measurable "real" variable (Figure 1), rather quality is the result of an arbitrary translation between a measurable reality and some subjective aspiration. Engineers showed a great ability to compute outcomes that (often) appeared to satisfy quality in practice. Also in this case, computations from syntactical data support "real" engineering decisions.

The Church-Turing thesis states:

> If something is computable on a discrete device, then it is also computable on a Turing Machine.

This implies that up until now, no one has been able to construct a discrete computing device for which an equivalent Turing Machine does not exist. Still, some computations that are computable on a Turing Machine in principle, may be too complex to compute on a practical device in time. Engineers showed great ability in constructing devices that autonomously compute "real" decisions as feedback control loops are ubiquitous. So, Von Leibnitz's

dream *often* became attainable after all. As an example, the feedback controller (C) shown in Figure 2 autonomously computes an input signal (U) to the process (P) that yields an output (Y). This computation depends on the error (W) between the output (Y) and the set point.

Still, the delimitations from section 2.1 remain unresolved implying that (*i*) engineering guidelines are occasionally improved by lessons learned from "real" disasters, or that (*ii*) the feedback control loop occasionally oscillates away from the set point. Where engineering computations fail, maintenance is often triggered. Maintenance is defined by:

> The combination of all technical and administ
> actions, including supervision actions, to retai
> restore an item's quality.

This definition paraphrases CEN (2019) and IEC (2015 maintenance is considered as a decision to act, wit intention to cause a quality effect. Figure 2 sho maintenance control loop that should correct the faults autonomous feedback control loop (Tinga et al., 2023) maintenance control loop is typically triggered b detection of a fault, i.e. an observation of some ano Fault isolation is the assignment of a specific fault labe assists in the choice of the recovery action. identification is an assessment of the (evolution in magnitude of the fault. Prognostics is an estimation remaining useful life. Finally, recovery is an actio causes quality. This maintenance control loop follows a Detection and Isolation (FDI) convention (Isermann, 2006).

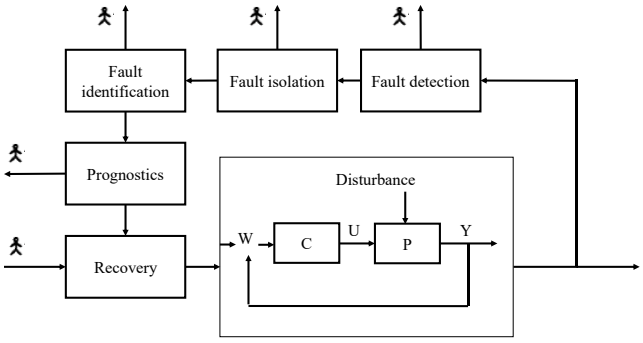


Figure 2. Autonomous control loop extended with maintenance control.

Although the maintenance control loop is thought to be human involved (as indicated by the person symbols in Figure 2), parts of it may still be computed. For example, the fault detection and the fault isolation may be computed before a human takes over. Then, this human may not need to troubleshoot the anomaly as this has been computed autonomously.

In conclusion, engineers have developed a great ability to compute "real" decisions and to construct devices that could similarly do so autonomously. Still, engineering computations occasionally fail which triggers human involved maintenance. Therefore, computing autonomous maintenance is challenging, but parts of the maintenance control loop may still be supported by computations. For that reason, the title of this paper refers to data driven maintenance rather than autonomous maintenance.

2.3. Maturity in data driven maintenance

Data maturity models are widely researched (Al-Sai et al., 2023) and applicable. Figure 3 shows a commonly adopted data maturity classification that includes monitoring,

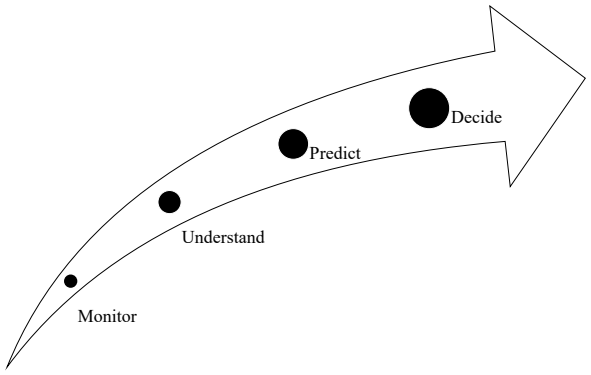


Figure 3. Data maturity model.

A comparison of the data maturity model in Figure 3 with the maintenance control loop in Figure 2 reveals that data maturity grows as more steps in the maintenance control loop are being computed, i.e. monitoring corresponds with fault detection, understanding with fault isolation & identification, predicting with prognosis and deciding with recovery.

Tiddens et al. (2023) observed a relation between an aspired maturity level and the required measurements. This paper intends to be more precise about this relationship by comparing two computations of fault detection and isolation that both provide a correct decision for a specific set of measurements. Still, these two computations will differ in attainable maturity as they translate to "reality" in a different way (Figure 1), i.e. the "real" causal implication of corresponding syntactical computations will be shown to differ. Then, the attainable maturity does not just rely on measurements, but also on a subjective translation.

2.4. Causality

This section will introduce two ways to address causality when computing a "real" decision (e.g. in the case study in the next section). In the experience based approach, a statistical association is computed and the causal assumptions are made separately. In the model based approach, the effect of setting a variable in an engineering (design) model of

equivalences is computed and the causality follows from the deterministic process of the computation itself.

An equivalence is symmetrical, reflexive, and transitive:

$$Y = aX + b \quad (1)$$

A causality is only transitive:

$$Y \leftarrow aX + b \quad (2)$$

In Eq. (1) swapping the terms around the equivalence symbol does not change the meaning of the expression. However, in Eq. (2) swapping the terms around the arrow changes the meaning of the expression from "X causes Y" to "Y causes X". It is important to realize that statistical associations retrieved from measurements are equivalences, but they do not imply causality.

Decisions rely on causality rather than on associations as a choice should bring about an effect that would not or less likely occur otherwise. To validate an individual decision, the effect of each choice would have to be observed whereas only the effect of the choice made is observable. Generally, the problem of observing the causal interactions in an individual experiment is that the counterfactuals remain unobservable. Therefore, the interventional distribution $Pr\big(Y|do(X)\big)$ may wildly differ from the observed distribution $Pr(Y|X)$.

Still, there are ways to strengthen a suspicion of causality across many experiments provided that the cause X in Eq. (2) sufficiently varies. Fisher (1935) proposed random assignment of treatments to eliminate the effect of unobserved confounders and he suggested that unobserved confounders could explain the measured association between smoking and lung cancer (Fisher, 1958). The latter beautifully illustrates the delicacy to use a measured association to support a decision to smoke. Structural Causal Modelling (SCM) proposed by Pearl (2009) also applies to non-experimental research constructs. SCM subsumes Structural Equations Modelling (Wright, 1934), and the Potential Outcomes Framework (Rubin, 2005). The experience based approach to the case study in section 4 will use SCM to specify the independence assumptions needed for a specific causal explanation of a computed statistical association.

Engineers typically use equivalence relations like bond graphs or finite element methods when designing a device. These equivalence relations are acausal, but the computation of their solution is a sequential process that introduces causality, i.e. if one variable in these equations has been set to a known value, the response of the other variables follows by computation. So, there is an intimate relationship between computing the solution of an engineering model and causality (Karnopp et al., 2012). The causal effect of a "real" decision to set one of these variables is similarly computable. The model based approach to the case study will use a bond graph to model the case study and the causality follows from the sequence in the computation itself.

In conclusion, this subsection showed that causality could be assigned after the computation of a statistical association and that causality is just inherent to the process of computing. Both notions of causality will be applied to the case study.

Now these four basic ingredients of data-driven maintenance decision making have been considered, the theoretical concepts will be converted to a practical application in the next two sections.

3. AUTONOMOUS FAULT DETECTION AND ISOLATION

This section will portray a typical construction of autonomous fault detection and isolation. Fault detection and isolation are the first "real" decisions in the maintenance control loop (Figure 2). A Fault Signature Matrix (FSM) will be used to assess the ability to detect or isolate faults. The rows in a FSM list the applicable faults (Table 1). A fault can be defined as an anomaly that precedes a failure (= nonconformity in quality). The columns in a FSM list the features (or symptoms) that indicate the faults (Table 1). The fields in a FSM indicate the relationship between the faults and the symptoms. A FSM could therefore support decisions to detect or to isolate faults (step 1 and 2 in Figure 2). For example, $Fault_0$ in Table 1 is detectable and isolable by the feature F_0. $Fault_1$ and $Fault_2$ are detectable but not isolable by the features F_1 and F_2, while $Fault_3$ is both detectable and isolable by these two features.

Table 1: Example of a FSM.

	F_0	F_1	F_2
$Fault_0$	1	0	0
$Fault_1$	0	1	1
$Fault_2$	0	1	1
$Fault_3$	0	0	1

An Experience Based (EB) and a Model Based (MB) approach to construct a FSM will illustrate two scenarios for the assignment of causality. It will become clear that an EB_FSM merely relates faults to associated symptoms and a causality assignment will require additional assumptions. For a MB_FSM, causality has already been settled in the process of its construction. The objective here is to explore the human involvement. The objective is *not* to review all existing approaches or to exhaustively review the computing of the fault detection and diagnostics. The presented FSM constructions just survey the essential steps to be taken in the

sim ify
inv

3.1.

Thi
will

Ste

The
pre
rele
tast
esta
con
(Ch
(RC
identifying the faults that may predict them.

The history of measurements will often be collected by non-experimental research which precludes control over the collection of *all* relevant fault states and operating regimes. By conceiving many fault states and operating regimes, the collection of the history of measurement may take too long (=complexity issue analogous to computational complexity). Moreover, faults are often a hidden variable. As already signaled by Tiddens et al. (2023), the history of measurements often delimits aspirations to compute fault detection and isolation.

Step 2: choose the features (EB_FSM columns).

The features of choice should be built from the history of measurements. A data scientist may generate an enormous amount of features from the library of signal features (Lu et al., 2023) while ignoring the choice of the faults. Burnham & Anderson (2002) already argued that even vague knowledge regarding related variables reduces the computational complexity of the model selection while avoiding spurious relations. Engineering guidelines may establish common sense about features (Isermann, 2011) that indicate a fault.

Step 3: select a model

Any regression or classification model may be considered to describe the data, but the shortest description is supposed to be the best one (Occam's razor). However, the shortest description of a data string is fundamentally incomputable (Solomonoff, 1964). Therefore, model selection remains rather arbitrary. Still, a suboptimal approximating model could support a dithering decision maker accepting some risk.

Step 4: explain the model

To explain the selected model, i.e. to identify which features strongly relate to a fault, some arbitrary feature importance test may be chosen. However, feature importance scores do not indicate causality, while a decision maker who does not only seek support in deciding *whether* to act, but also in *how* to act, requires causality. Section 2.4 mentioned that

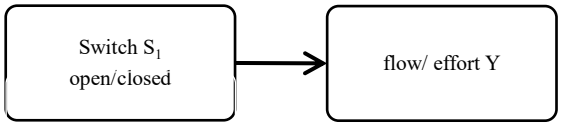


Figure 4. Example of a DAG.

Figure 4 is a directed acyclic graph (DAG) that specifies the causalities in a universe of the variables (S, Y). By Bayesian Network Factorization, the joint probability distribution $Pr(S, Y)$ follows from the DAG in Figure 4:

$$Pr(S, Y) = Pr(S)Pr(Y|S) \quad (3)$$

Eq. (3) specifies the potentially observable association to identify a causality provided that the DAG is true. For example, the causality $Pr\big(Y|do(S)\big)$ is identifiable by the potentially observable association $Pr(Y|S)$, provided that Figure 4 is true. The DAG may be highly controversial, but it is explicit at least (Pearl, 2009).

3.2. Model based fault signature matrices

This section will outline the construction of a MB_FSM that will be used in the case study.

Step 1: construct an engineering model

A device does not come from some natural phenomenon, it is the result of a deliberate design. Engineers typically compute their designs using the laws of physics. These laws of physics hold under idealized conditions and they should adequately approximate the "real" conditions. These approximations are usually reflected in engineering guidelines that prescribe safety margins. Laws of physics and engineering guidelines are arbitrary in principle as they are occasionally updated, but they generally reflect a very high degree of common sense.

Step 2: choose the faults (MB_FSM rows)

Faults should be phrased in terms of drifts in parameters in the engineering model. If other faults (beyond the parameters in the model) should be detected or isolated, the engineering model needs extension or an additional EB_FSM may be needed.

Step 3: choose the ARRs (MB_FSM columns)

From an engineering model of n equations the values of n variables are computable. As (some of) these variables are measured, less equations are needed which enables the formulation of Analytical Redundancy Relations (ARR). An ARR is an equivalence consisting of measurements and parameters from the engineering model. An ARR detects faults that have been defined as parameter drifts, and thus acts as feature or symptom in the FSM.

Step 4: construct the MB_FSM

The faults (MB_FSM rows) have been defined at step 2. The *ARR*s have been defined at step 3, and the fields trivially follow from the presence of the parameters in the *ARR*s. Therefore, the construction of the MB_FSM is autonomously computable from the previous steps.

The *ARR* s are acausal equivalence relations. However, computing the solution of the *ARR*s involves a sequential process where the values of the *ARR*s follow from their variable and parameter values. Similarly, a "real" decision to set a variable or a parameter to a specific value causes the corresponding *ARR*s to change. As a fault in step 2 has been defined as a drift in some *ARR* parameter, this fault causes the *ARR* s to change within the universe of idealized conditions of the engineering (design) model from step 1.

4. CASE STUDY

This section will demonstrate fault detection and isolation by constructing an EB_FSM and a MB_FSM in an iconic case study of a linear time invariant system under feed-forward

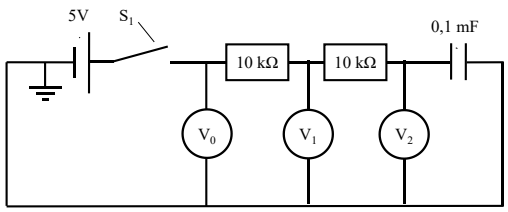


Figure 5. The RRC circuit.

A pulse signal with a period of 20 seconds will trigger the switch S_1. The lines in Figure 6 show the computed evolution of the voltages and the dots show the measured evolution of the voltages for a normal (healthy) state of the circuit.

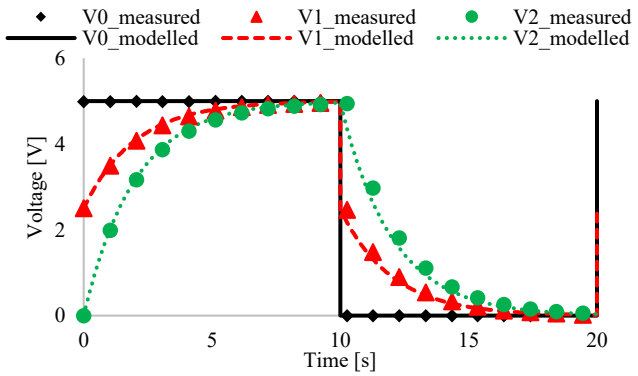


Figure 6. Evolution of the computed and the measured voltages at the healthy state.

Figure 6 confirms that engineers are highly capable of deciding about the "real" behavior of the RRC circuit by computation. Occasionally, the "real" measurements may drift away from the engineering computation which could trigger maintenance. In this case study, two fault treatments have been applied:

1. A decreased resistance R_0 that is in between the voltages V_0, V_1 in Figure 5.
2. An increased capacitance.

Fault detection and isolation would have been trivial if the resistance and the capacitance were directly observable. It is only due to the experimental setup of this case study that the presence and absence of the faults was certain. Therefore, fault labels in Figure 7 and Figure 8 just followed from a known experimental intervention.

Figure 7 shows that in the faulty state (reduced resistance) the measured voltages respond faster to the switch than predicted by the engineering computation (for the healthy state).

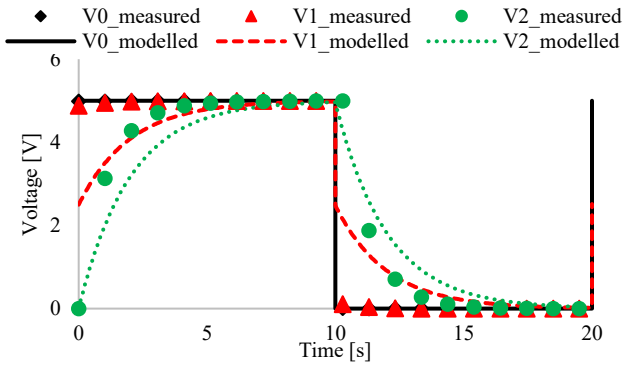


Figure 7. Evolution of the computed and the measured voltages at a decreased resistance R_0.

Figure 8 shows that the measured voltages respond slower to the switch at an increased capacitance than predicted by the engineering computation.

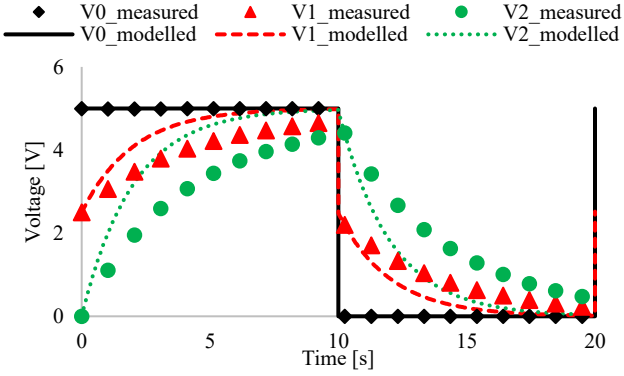


Figure 8. Evolution of the computed and the measured voltages at an increased capacitance.

Note that the operating regime of a pulse signal h influences Figure 6, Figure 7 and Figure 8 as the RRC c is known to operate as a low-pass filter.

Section 4.1 and section 4.2 will explain the construction EB_FSM and a MB_FSM respectively. The focus will the possible obstructions (section 2.1) in computing detection and isolation and not on a quest for the op computation.

4.1. Application of EB_FSM

Let the faults (EB_FSM rows) be a reduced resistance a increased capacitance. Let the features (EB_FSM colu be the *measured* voltages V_0, V_1, V_2, the switch positic and the time T from Figure 6, Figure 7, and Figure 8. that the lines in the three plots are the predictions of an engineering (design) model that should be ignored here.

Let the fields of the EB_FSM be the permutation importance scores of an arbitrary random forest classification. The permutation importance indicates the mean Gini impurity loss of the random forest classification after random resampling of a feature. Note that the EB_FSM fields do not only rely on aforementioned choices, but also on the history of measurements in Figure 6, Figure 7, and Figure 8.

Then, the EB_FSM is given in Table 2.

Table 2: EB_FSM of the case study.

	V_0	V_1	V_2	T	S_1
Resistance R_0 ↓	0,00	0,30	0,06	0,04	0,00
Capacitance ↑	0,04	0,18	0,12	0,16	0,00

Table 2 shows that the voltage V_1 entailed unique information about a decreased resistance as random resampling strongly affects the mean Gini impurity loss of the random forest classification. Similarly, the voltages V_1, V_2, and the time T entailed unique information about an increased capacitance.

The EB_FSM may be used to reduce the complexity of the model selection as Table 2 implies that the random forest classification could still detect both faults when the switch position S_1 is omitted from the history of measurements.

As this paper is not about an improved model selection, details about the arbitrarily selected model will be omitted. It has just been verified that the model of choice correctly predicted all instances in a validation set comprising the same faults that occurred during the same operating regime. So, fault detection and fault isolation (Figure 2) is possible for this specific validation set.

Let the DAG in Figure 9 apply to the EB_FSM (Table 2). This DAG asserts that changes in the resistance R_0, in the capacitance C, or in the switch S_1 cause some hidden flow

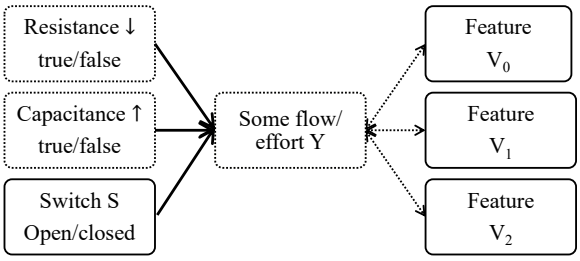


Figure 9. DAG with indicators.

It has been presumed that the switch S_1 in the DAG (Figure 9) does not cause the faults and the EB_FSM confirms that the switch S_1 neither associates with the faults. Similarly, it has been presumed that the time T does not cause the faults (not in DAG) but the EB_FSM shows that the time T still associates with the faults. Still, section 2.4 already mentioned that observed associations (in the EB_FSM) are not compelling for a DAG. A DAG merely specifies the independence assumptions (omitted arrows) of a specific causal explanation for the EB_FSM.

Section 3.1 mentioned that a decision regarding the fault detection or isolation may be incomputable because it is fundamentally incomputable, it is too complex, or it is subject to philosophical controversy. In this case study, the latter prevailed as the DAG is merely postulated afterwards. Therefore, a compelling causal explanation of the computed fault detection and isolation is lacking. In other words, the causality is philosophically controversial. Common sense reflected in engineering guidelines (section 3.1) may alleviate this controversy. The effects of this controversy are:

- Fault detection and isolation beyond the history of measurements (training set) is risky.
- The applicability of the fault detection and isolation is unknown, i.e. it worked for a specific validation set, but it is unknown whether it will work at an unprecedented operating regime.
- The features (like the time T) do not necessarily indicate the magnitude of the fault.

Finally, the fault detection and isolation relied on the arbitrary choice of the classification model, and the feature importance score. Different results might have been obtained had other choices been made.

4.2. Application of MB_FSM

In advance of constructing a MB_FSM, an engineering (design) model will be posited. Let the case study be represented by the Hybrid Bond Graph (HBG) in Figure 10.

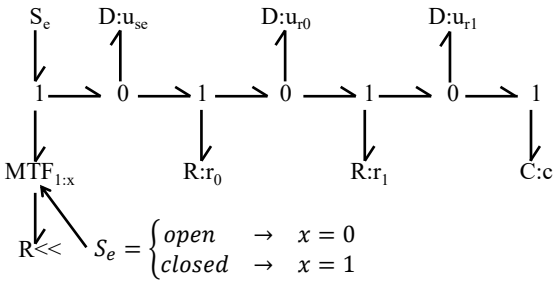


Figure 10. Hybrid Bond Graph of the case study.

The switch has been modelled by a modulated transformer (MTF) as proposed by Borutzky (2012). Figure 10 shows four elements that convert power. As power is the product of an effort variable and a flow variable, the engineering model (Table 3) consists of eight variables and eight constitutive equations that follow from Ohm's Law and Kirchhoff's Law. In this case study, the effort of the source $u_{S_e} = V_0$, and the effort of the resistances $u_{R_0} = V_0 - V_1$, $u_{R_1} = V_1 - V_2$ have been measured which makes three of the equations in Table 3 redundant.

Table 3: Engineering (design) model for case study.

i	$u_{S_e} = 5 \times x \,;\, x \in \{0,1\}$
ii	$0 = u_{R_0} - 10^4 \times i_{R_0}$
iii	$0 = u_{R_1} - 10^4 \times i_{R_1}$
iv	$0 = u_C - 10^4 \times \int i_C(t)dt$
v	$0 = u_{R_0} + u_{R_1} + u_C - u_{S_e}$
vi	$0 = i_{R_0} - i_{R_1}$
vii	$0 = i_{R_0} - i_C$
viii	$0 = i_{R_0} - i_{S_e}$

Let's now construct an MB_FSM of the case study using this engineering model. Let the faults (MB_FSM rows) be a drift in the resistance R_0 and a drift in the capacitance C. As a drift may include an increase as well as a decrease, these fault definitions are more generic than the ones in Figure 7 and Figure 8. Note that the history of measurements (Figure 6, Figure 7 and Figure 8) is not needed for the construction of a MB_FSM.

Let the features (MB_FSM columns) be defined by the ARRs that follow from the measured variables in the engineering model (Borutzky, 2021), (Samantaray et al., 2006).

The ARR_1 is given by:

$$0 = \frac{V_0 - V_1}{R_0} - \frac{V_1 - V_2}{R_1} \tag{4}$$

The ARR_1 follows from (ii), (iii) and (vi) in Table 3, and the voltages V_0, V_1, and V_2.

The ARR_2 is given by:

$$0 = V_0 - V_2 - (V_{0x} - V_{2x}) \times e^{-\frac{(T-x)\times C}{R_0+R_1}} \tag{5}$$

In Eq. (5), V_{0x}, V_{2x} represent the voltages at the time of the last switch transition. The ARR_2 follows from (iv) and (v) in Table 3, the evolution in u_{S_e}, and the measurements V_0, V_2, T.

Let the fields of the MB_FSM be given as shown in Table 4, revealing an indicator function on the presence of the drifting parameters in the ARRs..

Table 4: MB_FSM of the case study.

	ARR_1	ARR_2
Drift in R_0	1	1
Drift in C	0	1

Now, the MB_FSM could be used to evaluate the same validation set as the one used for the EB_FSM. Figure 11 confirms that both ARRs drift away from zero at a decreased resistance as predicted in the MB_FSM (Table 4).

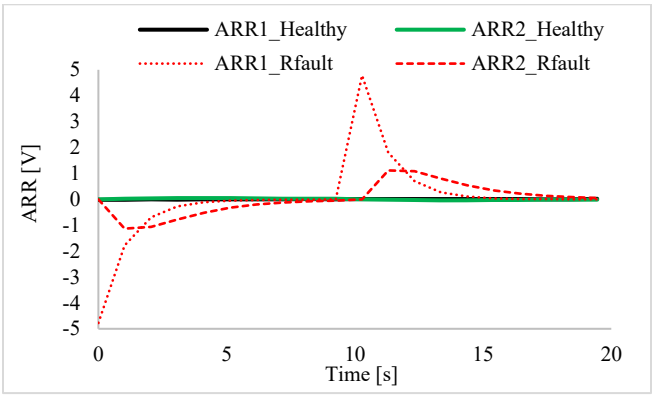


Figure 11. Measured ARRs at a decreased resistance.

Figure 12 confirms that only ARR_2 drifts away from zero at an increased capacitance as predicted in the MB_FSM (Table 4). By choosing a threshold value for the ARRs, the fault detection is autonomously computable. Figure 11 and Figure 12 show that the ARR s can only detect faults as the components in the RRC circuit exchange power shortly after a transition of the switch. The applicability of the fault detection and isolation under various switching regimes is straightforwardly assessable without any history of measurements.

Eq. (4) and Eq. (5) specify the value of the ARR at a given magnitude of the drift in R_0 or C, i.e. a decision regarding the fault identification (i.e. severity of the fault) is partially computable. An autonomously computable fault identification implies a higher maturity in data driven maintenance support (Figure 2) than just isolating the fault. Moreover, the impact of the precision of the measurements is assessable at the stage of design. The precision of the measurements is important to define appropriate threshold values on the ARRs.

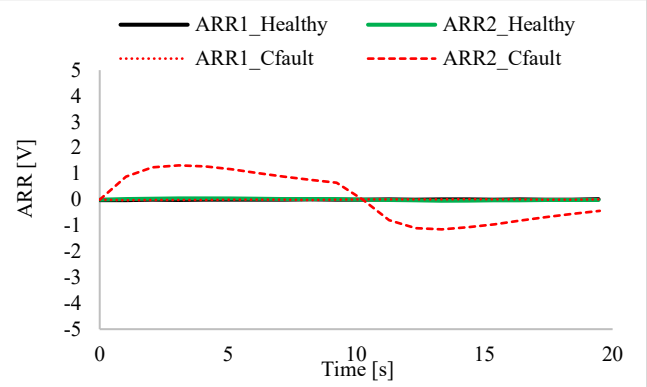


Figure 12. Measured ARRs at an increased capacitance.

Section 2.1 mentioned that a decision regarding the fault detection or isolation may be incomputable because it is fundamentally incomputable, it is too complex, or it is philosophically controversial. In this case study, the latter prevailed as the engineering model is not pertinently true. The fault detection and isolation relied on the applicability of the idealized conditions of the laws of physics that underlie the engineering model. Typically, physical laws are rather robust against changes in these conditions. Still, unmeasured operating conditions may become problematic. For example, large but unrecorded temperature fluctuations may trouble Ohm's Law and consequently the fault detection and isolation (MIL-HDBK-217F, 1991).

If the engineering (design) model were to be true, the MB approach would have resolved the concerns of the EB approach:

- Fault detection and isolation beyond the history of measurements (training set) is decidable. The MB_FSM can even be constructed at the stage of design (without any training set at all).
- The applicability of the fault detection and isolation to work is known. For example, it is known that the fault detection and isolation only works as power is being exchanged.
- The ARRs indicate the magnitude of the fault. Therefore, the attainable maturity in data driven maintenance is potentially higher.

Finally, section 3.2 mentioned that the engineering (design) model may just be incapable to detect or isolate a particular fault. As aspirations should meet capabilities, the engineering (design) model may need adjustments for the purpose of data driven maintenance.

5. Discussion

This section will reflect on the case study. Section 5.1 will discuss the impact on the computability of "real" decisions, section 5.2 will discuss the impact on the maturity in data driven maintenance, and section 5.3 will discuss some practical implications.

5.1. Impact on computing "real" decisions

Section 2.1 mentioned that a decision may be incomputable because it (*i*) is fundamentally incomputable, it (*ii*) is too complex, or it (*iii*) is philosophically controversial.

In this simple case study, the philosophical concerns appeared predominant as the translation between a syntactical computation and a "real" decision required arbitrary human involvement to choose:

- The faults (EB, MB);
- The measurements/ features (EB/MB);
- A classification model (EB/MB);
- A feature importance score (EB);
- A causal explanation (EB);
- An engineering (design) model (MB).

The engineering profession established a high degree of common sense regarding this translation by formulating laws of physics and guidelines. This common sense lacks the solidity of a mathematical proof, and it has been subject to occasional improvement, but it has shown to be effective due to the wide application of engineered devices. Section 2.2 stated that where engineers fail to compute "real" decisions, a human involved maintenance control loop is typically triggered. Still, parts of the maintenance control loop may be computed as shown in the case study. Cases where the computing of "real" decisions is challenging, are also expected to be of high interest to scientists.

In the simple case study, complexity was not an issue. Still, complexity plays a role in other cases. For the EB approach, the inference of a high dimensional model from a large history of measurements may require excessive computing time. Section 3.1 stated that complexity may impede the collection of a history of measurements that includes all relevant system states. Particularly under a non-experimental research construct, the required time is uncontrolled. For the MB approach, the solving of a high dimensional engineering (design) model may similarly bump into complexity concerns.

Fundamental incomputability precluded the selection of a true EB_FSM model (section 3.1). Similarly, the truth of the engineering model (Table 3) was ultimately an incomputable postulate. Fundamental incomputability is also an issue in cases of software faults as there cannot exist a computing device that separates looping software from software that halts in the general case. If this computing device only had to separate software of some fixed number of input symbols, the computation rapidly becomes too complex to solve in time (Rado, 1962).

5.2. Impact on maturity

Growth in the data maturity model (Figure 3) coincided with the flow of the consecutive decisions in the maintenance control loop (Figure 2). This paper confirms that the computation of fault detection and isolation should be settled before addressing the computation of decisions further downstream the maintenance control loop. Similarly, maturity growth in data driven maintenance should start with computing fault detection and isolation.

In the specific validation set of the case study, the EB approach and the MB approach were exchangeable in terms of missed and false alarms. Still, a decision maker should not be indifferent towards the approach because (*i*) causality is assigned differently, and (*ii*) the meaning of the features differs. Using the EB approach, causality was assigned afterwards using some arbitrary DAG and the features just described the state of the RRC circuit. Using the MB approach, causality was inherent in the solving of the engineering (design) model and the ARRs represented the magnitude of the fault. The latter is part of fault identification (Figure 2) which corresponds with a higher maturity in data driven maintenance.

5.3. Practical impact

The case study revealed that the "real" causal implications of some syntactical computation matter for the attainable maturity in data driven maintenance. In the cases study, both the EB and the MB approach appeared to be not entirely compelling for causality. Still, some references to engineering guidelines were given to alleviate potential controversy. Section 3.1 referred to some engineering guidelines for (*i*) the most relevant faults of specific devices and for (*ii*) typical features to detect these faults. Section 3.2 referred to some engineering guidelines to establish common sense regarding the margins between the computed strength and the "real" strength.

For this iconic case study, the construction of a MB_FSM was easy but for a more realistic case study, the construction of a MB_FSM could become complex. Typically, the knowledge of the engineering models is scattered over various agents who may be unwilling to share them. Consequently, much effort may be wasted on reconstructing design models that are in principle already available. Life cycle modelling as proposed in ISO (2014) is a precondition to apply a MB_FSM efficiently in practice.

The EB approach and the MB approach do not compete as one may also consider a hybrid FSM that adds the ARRs to an EB_FSM. The EB approach that decides on associated symptoms may be an appreciable resort in the absence of a causal explanation. The MB approach demonstrated the potential of a more mature data driven maintenance under idealized conditions.

6. CONCLUSION

This paper argued that some decision problems cannot be solved by any autonomous computation and that maintenance decisions are prone to be computationally challenging. A maturity framework has been proposed that specifies the decisions in a maintenance control loop, and connects these to the aspects of human interpretation, computability and causality. An application of the lowest maturity level to an iconic case study showed that decision makers should not be indifferent to (two) models that provide equal decisions on a validation set in terms of missed and false alarms. Access to a true engineering (design) model allows achieving a higher maturity level in data driven maintenance but it has been observed that a true model cannot be computed from only a history of measurements. Where logic cannot decide, the common sense reflected in engineering guidelines provides a resort at an acceptable risk.

ACKNOWLEDGEMENTS

This research is part of the European Digital Naval Foundation (EDINAF) Project, a project that has received funding from the European Defence Fund (EDF) under grant agreement 101103273 - EDINAF - EDF-2021-NAVAL-R-2. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission (the granting authority). Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Al-Sai, Z. A., Husin, M. H., Syed-Mohamad, S. M., Abdullah, R., Zitar, R. A., Abualigah, L., & Gandomi, A. H. (2023). Big Data Maturity Assessment Models: A Systematic Literature Review. *Big Data and Cognitive Computing*, *7*(2), 1–28. https://doi.org/10.3390/BDCC7010002

Borutzky, W. (2012). Bond-graph-based fault detection and isolation for hybrid system models. *Proceedings of the Institution of Mechanical Engineers. Part I: Journal of Systems and Control Engineering*, *226*(6), 742–760. https://doi.org/10.1177/0959651812440665

Borutzky, W. (2021). Fault Diagnosis. In *Bond Graph Modelling for Control, Fault Diagnosis and Failure*

Prognosis (pp. 51–130). Springer International Publishing. https://doi.org/10.1007/978-3-030-60967-2_3

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference* (2nd ed.). Springer New York. https://doi.org/10.1007/B97636

CEN. (2007). *EN 1990-1999: Eurocodes for buildings*. European Committee for Standardisation.

CEN. (2019). *EN 13306: Maintenance terminology*. European Committee for Standardisation.

Chandler, G., Denson, W. K., Rossi, M. J., & Wanner, R. (1991). *Failure Mode/Mechanism Distributions*. Reliability Analysis Center.

Church, A. (1936). An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, *58*(2), 345. https://doi.org/10.2307/2371045

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Fisher, R. A. (1958). Lung cancer and cigarettes? *Nature*, *182*(4628), 108. https://doi.org/10.1038/182108a0

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte Für Mathematik Und Physik*, *38*(1), 173–198. https://doi.org/10.1007/BF01700692/METRICS

Hoffman, D. D. (2019). *The case against reality: why evolution hid the truth from our eyes*. W.W. Norton & Company.

IACS. (2024). *Blue book.* International Association of Classification Societies.

IEC. (2015). *IEC 60050: International Electrotechnical Vocabulary (IEV) - Part 192: Dependability*. International Electrotechnical Commission.

Isermann, R. (2006). Fault-diagnosis systems: An introduction from fault detection to fault tolerance. In *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-30368-5/COVER

Isermann, R. (2011). *Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-tolerant Systems*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12767-0

ISO. (2014). *ISO 55000: Asset management — Overview, principles and terminology*. International Organisation for Standardisation.

ISO. (2015). *ISO 9000: Quality management systems — Fundamentals and vocabulary*. International Organisation for Standardisation.

ISO. (2016). *ISO 14224: Petroleum, petrochemical and natural gas industries: Collection and exchange of reliability and maintenance data for equipment*. International Organisation for Standardisation.

Karnopp, D., Margolis, D., & Rosenberg, R. (2012). *System Dynamics: Modeling, Simulation, and Control of Mechatronic Systems* (5th ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118152812

Lu, S., Lu, J., An, K., Wang, X., & He, Q. (2023). Edge Computing on IoT for Machine Signal Processing and Fault Diagnosis: A Review. *IEEE Internet of Things Journal*, *10*(13), 11093–11116. https://doi.org/10.1109/JIOT.2023.3239944

MIL-HDBK-217F. (1991). *MIL-HDBK-217F: Reliability prediction of electronic equipment*. U.S. Department of Defense.

OREDA. (2002). *OREDA: offshore reliability data handbook* (4th ed.). OREDA Participants.

Pearl, J. (2009). Causality: Models, reasoning, and inference, second edition. In *Causality: Models, Reasoning, and Inference, Second Edition* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Rado, T. (1962). On Non-Computable Functions. *Bell System Technical Journal*, *41*(3), 877–884. https://doi.org/10.1002/J.1538-7305.1962.TB00480.X

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Samantaray, A. K., Medjaher, K., Ould Bouamama, B., Staroswiecki, M., & Dauphin-Tanguy, G. (2006). Diagnostic bond graphs for online fault detection and isolation. *Simulation Modelling Practice and Theory*, *14*(3), 237–262. https://doi.org/10.1016/J.SIMPAT.2005.05.003

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*(1), 1–22. https://doi.org/10.1016/S0019-9958(64)90223-2

Tiddens, W., Braaksma, J., & Tinga, T. (2023). Decision framework for predictive maintenance method selection. *Applied Sciences*, *13*(3). https://doi.org/10.3390/APP13032021

Tinga, T., Homborg, A. M., & Rijsdijk, C. (2023). Data-driven maintenance of military systems: Potential and challenges. In P. B. M. J. Pijpers, M. Voskuijl, & R. Beeres (Eds.), *Towards a data-driven military. A multi-disciplinary perspective* (pp. 73–96). Leiden University Press. https://doi.org/10.24415/9789087284084

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. https://doi.org/10.1112/PLMS/S2-42.1.230

Wright, S. (1934). The Method of Path Coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215. http://www.jstor.org/stable/2957502

A Novel Approach for Evaluating Datasets Similarities Based on Analytical Hierarchy Process in the Industrial PHM Context

Mohamed Aziz Zaghdoudi[1], Christophe Varnier[2], Sonia Hajri-Gabouj[3], and Noureddine Zerhouni[4]

[1,2,4] *SUPMICROTECH, CNRS, institut FEMTO-ST, F-25000 Besançon, France*
mohamed.zaghdoudi@femto-st.fr
christophe.varnier@ens2m.fr
noureddine.zerhouni@ens2m.fr

[1,3] *LISI, Institut National des Sciences Appliquées et de Technologie, Université de Carthage, Centre Urbain Nord, BP 676, 1080, Tunis, Tunisia*
sonia.hajri@insat.ucar.tn

ABSTRACT

In prognostics and health management (PHM), data-driven approaches are crucial for performing prognostics based on historical data, relying on the analysis of extensive datasets to identify patterns and relationships that contribute to predicting or optimizing variables. However, their efficiency is contingent upon the availability of large, high-quality datasets tailored to the specific task at hand.

Yet, real-world applications frequently face challenges as data may not always be readily available due to limitations in data acquisition systems or confidentiality concerns. Paradoxically, the contemporary era witnesses an unprecedented surge in the availability of online databases across various fields. These databases offer a plethora of data that can be harnessed to develop, prototype, and test PHM solutions.

This study endeavors to introduce an innovative approach for assessing the similarity between datasets, specifically tailored for prognostic and health management applications. The objective is to empower the development of PHM solutions for predefined systems without relying on data generated from the system itself, but rather by leveraging analogous datasets. To quantify the similarity between different datasets, we propose a set of criteria and sub-criteria based on the characteristics of datasets. Subsequently, the analytic hierarchy process (AHP), a well-established multi-criteria decision-making approach, is employed to systematically compare the importance of criteria and sub-criteria for each elementary process within the PHM cycle. This dynamic process considers the varying importance of criteria across different phases, acknowledging that a criterion may not be uniformly significant for all elementary processes. The evaluation of dataset similarity incorporates the proposed criteria and sub-criteria, utilizing a fundamental scale of importance intensity and weights assigned through AHP. This holistic approach yields a comprehensive similarity score, enabling a nuanced understanding of dataset compatibility.

To exemplify the efficiency of our proposed approach, we applied it to a practical case study. The study involves assessing the similarity between a run-to-stop database of mechanical bearings and a set of online databases dedicated to the same application. Our solution facilitated the identification of criteria pertinent to the case study, the determination of criterion weights, and ultimately, the calculation of a similarity score for each database. This process proved instrumental in selecting the most similar database, showcasing the practical utility of our proposed approach in real-world PHM scenarios.

1. INTRODUCTION

Prognostics and Health Management (PHM) is an engineering and research field that aims to study fielded systems conditions, predict their possible failures, and take appropriate actions to mitigate those malfunctions effects (Bougacha, Varnier, & Zerhouni, 2022). In this context, data-driven approaches are being increasingly used to convert historical data into models that accurately represent the physical systems' degradation behavior (Tobon-Mejia, Medjaher, Zerhouni, & Tripot, 2012). To perform efficiently, those approaches require the presence of extensive datasets, adhering to established data quality standards, and accurately reflecting the characteristics of the system under study. However, for real systems, data collection is a complicated process that requires setting up sometimes costly acquisition devices, overcoming confidentiality issues, and selecting the characteristics of the data to be collected (data format, relevant variables, data quality

requirements...). This has led to a problem of insufficient amount of data for some PHM applications and uncertainty regarding the characteristics of the data to be collected.
Conversely, the current era is experiencing a proliferation in both the quantity and diversity of online databases, with approximately 31 million databases accessible on the Internet as of August 2020 (Benjelloun, Chen, & Noy, 2020). These publicly accessible datasets span a broad spectrum of domains, encompassing around 4600 domains in August 2020 (Benjelloun et al., 2020), and are amenable to adaptation for analogous problem-solving scenarios.
This theme has motivated this research work. We are interested in finding an approach for datasets similarity evaluation that makes it possible to find, among freely accessible datasets, the most similar dataset to a sample of data from a system studied in order to overcome the problem of lack of data for PHM applications.
In pursuit of this goal, we have introduced a set of criteria grounded in data characteristics to assess the similarity between datasets. Subsequently, we presented a methodology employing the Analytical Hierarchy Process (AHP), a widely recognized multi-criteria decision-making technique. This methodology serves to determine criteria weights and evaluate datasets similarity on the base of those criteria.
The remainder of this paper is organized into four sections. Section 2 summarizes previous works related to data insufficiency, data characterization, and the AHP technique. Section 3 describes the proposed methodology. Section 4 presents an illustrative case study evaluating the similarity between different bearing datasets. In section 5, a reliability evaluation approach is proposed to assess the consistency of the results. Finally, section 6 summarizes the main findings and outlines future directions for research.

2. Related Work

2.1. Solving the data insufficiency problem

The data insufficiency problem was the subject of several research works. Indeed, (Guo, Lei, Xing, Yan, & Li, 2018) require the existence of two conditions for the success of machine diagnosis data-driven intelligent approaches : Labeled data containing fault information is available and training and test data are drawn from the same probability distribution. However, for some systems, it is difficult to obtain massive labeled data (Guo et al., 2018).
One of the solutions proposed in the literature is Transfer Learning. It is defined as follows: Given a source domain D_S with a corresponding source task T_S and a target domain D_T with a corresponding task T_T , transfer learning is the process of improving the target predictive function $f_T(\cdot)$ using related information from D_S and T_S , where $D_S \neq D_T$ or $T_S \neq T_T$ (Weiss, Khoshgoftaar, & Wang, 2016).
The transfer learning approach has been applied to several industrial systems. (Wen, Gao, & Li, 2017) applied deep transfer learning method for fault diagnosis in a big data environment. Their approach was tested on a Case Western Reserve University bearing dataset (Smith & Randall, 2015). (Shao, McAleer, Yan, & Baldi, 2018) developed a deep transfer learning framework for mechanical fault diagnosis and classification, and created a repository of several reference datasets.
Despite its ability to solve the data gap problem, the transfer learning technique requires that the source and target data are similar and of the same distribution.
Another widely used approach is data augmentation. This technique consists in increasing the amount of training data by using the information contained within it (Perez & Wang, 2017).
Various data augmentation techniques have been applied to specific problems. The main techniques fall under the category of data warping, which is an approach to directly augment the input data to the model in the data space. This technique has been applied for several industrial applications and on various types of data. (Li, Zhang, Ding, & Sun, 2020) employed it for fault diagnosis of rotating machines. They applied 5 techniques for data augmentation in the form of digital signals, namely, Gaussian noise, masking noise, signal translation, amplitude shift, and time stretching.
Moreover, this technique is widely used with image data. As an example, we cite the work of (Wang, Yang, Jiang, & Fan, 2020) on image augmentation for crack detection using 9 different techniques.
Certainly, the data augmentation technique is useful to overcome the problem of lack of data for different applications and data types. However, this approach requires the existence of a minimal amount of data to be augmented.
On the other hand, other alternatives are used by researchers and industrialists to generate artificial data, such as physical model-based simulation (Saxena, Goebel, Simon, & Eklund, 2008) or test bench fabrication (Nectoux et al., 2012).

2.2. Analytical Hierarchy Process

The Analytical Hierarchy Process (AHP) was developed by Saaty in the 1970s (Saaty, 1980). This method, used in many fields related to multiple criteria decision-making (MCDM) is considered one of the most useful decision-making techniques (Ahmadi, Arasteh Khouy, Kumar, & Schunnesson, 2009). It's a methodology for relative measurement (Brunelli, 2014) where the focus is on proportions between some quantities rather than their exact measurement.
In AHP, The problem is divided into a hierarchy of qualitative and quantitative criteria, and then, using experience, the degree of relative importance is deducted. According to (Nydick & Hill, 1992), the AHP method is based on 4 steps :

1. Problem structuring
2. Data collection and measurement
3. Normalized weights determination

4. Application and problem-solution-finding

The Analytical Hierarchy Process has been used in several industrial applications to make decisions in different areas. (Cabrita & Frade, 2016) proposed an AHP-based solution to the supplier selection problem using fourteen different criteria. (Ren & Lützen, 2015) used AHP for fuel evaluation and selection under nine criteria for emission reduction from shipping. (Kilic, Zaim, & Delen, 2014) evaluated and selected the best ERP system using an AHP-based solution to solve this MCDM problem.
Hence, the analytical hierarchy process can be considered as a strong decision-making tool that can be used to evaluate and select the best action/alternative in multiple criteria decision-making problems.

2.3. Data Characterization

Databases similarity assessment first requires the establishment of data characterization criteria. Several previous works have addressed the issue of database characterization. However, the definitions and criteria proposed differ from one work to another, and the research has not resulted in unified criteria.
In this context, (Alelyani, Liu, & Wang, 2011) proposed 4 characteristics and studied their effects on feature selection stability. The proposed characteristics are the number of samples, features and classes, and the data distribution. (Bhatt, Thakkar, & Ganatra, 2012) divided thirteen characterization criteria into 2 different groups: phenotype characteristics dealing with entropy and the noise-signal ratio, and characteristics concerning the genotype of a dataset, divided into 2 categories:

- Simple Characteristics concern the attributes and instances' numbers
- Statistical Characteristics that deal with the statistical aspect of data.

(Oreski, Oreski, & Klicek, 2017) characterized data by 11 characteristics in 5 different groups, consisting mainly of standard, data sparsity, statistical, information-theoretic, and noise measures.
On the other hand, data quality has emerged as a fundamental notion for characterizing data. (Strong, Lee, & Wang, 1997) have defined high-quality data as data that is suitable for data consumers. Thus, we can conclude that data with different degrees of quality will lead to different results. (Redman, 1997) proposed four data quality characteristics most studied in the literature: accuracy, consistency, completeness, and timeliness. (Omri, Al Masry, Mairot, Giampiccolo, & Zerhouni, 2021) suggest that for PHM applications, data quality is characterized by volume, accuracy and completeness.

3. Proposed Approach

The proposed methodology (Fig. 1) is composed of four different phases. The first phase includes the proposal of similarity criteria and sub-criteria. The second phase is linked to the PHM cycle and the processes that make it up. The third phase details the criteria and sub-criteria weights calculation using AHP technique. The final phase is dedicated to decision-making using the established methodology.

3.1. Problem modeling / Criteria setting

The first step consists of proposing similarity criteria according to which the similarity will be evaluated. This step is also called 'Problem modeling' for AHP applications (Ishizaka & Labib, 2011). In fact, it is recommended to structure the criteria in a hierarchical structure to be able to focus on their importance when assigning their weights (Ishizaka & Labib, 2011). A structure of sub-criteria assembled in clusters (criteria) helps describe the problem more conveniently and reduces bias (Ishizaka, 2004).
To define criteria that are in line with this problem, we mainly rely on the data characterization criteria proposed in the literature. In (Table 1), a non-inclusive list of 17 sub-criteria divided into four criteria is proposed to evaluate the similarity between databases. These criteria can be used fully or partially, depending on the application or case study under consideration.
In addition to the attributes outlined in existing literature, we have introduced two supplementary sub-criteria, namely 'Data extension' and 'Data format.' Specifically, within the context of a given system and application, data representing the system state may manifest in various types and formats, such as images, signals, or tabular data. Disparities in data format and extension necessitate distinct characterizations and treatments.
Furthermore, our research proposes a novel set of application-related criteria, consisting of two sub-criteria. These criteria aim to evaluate the domain (e.g., manufacturing, medical, transportation) of the system depicted in the dataset, along with discerning the data source—whether it originates from a real-world application, a simulation, or a test bench.

3.2. PHM cycle modeling

In order to assign weights to each similarity criterion, we propose to, firstly, divide the studied PHM cycle into elementary processes. In fact, the PHM cycle is composed of seven elementary processes according to (Omri, Al Masry, Mairot, Giampiccolo, & Zerhouni, 2020), namely data acquisition, data processing, data assessment, diagnostic, prognostics, decision support, and HMI. From data acquisition to decision support and HMI, the importance of each of the established criteria depends on the process.

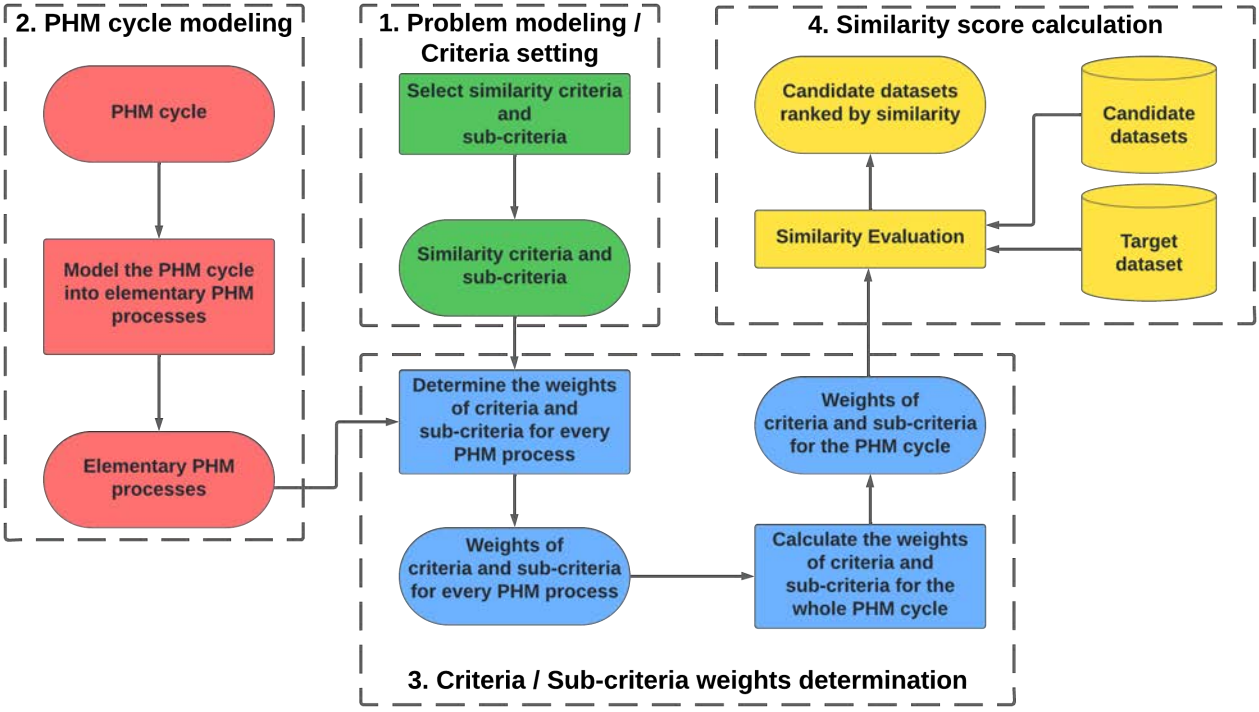


Figure 1. AHP based approach for datasets similarity evaluation

Table 1. Data similarity criteria

Criteria	Sub-criteria
Standard criteria	1. Number of attributes (Alelyani et al., 2011; Bhatt et al., 2012; Oreski et al., 2017) 2. Number of instances (Alelyani et al., 2011; Bhatt et al., 2012; Oreski et al., 2017) 3. Number of classes (Alelyani et al., 2011; Bhatt et al., 2012; Oreski et al., 2017) 4. Number of binary features (Bhatt et al., 2012) 5. Data format 6. Data extension
Statistical criteria	7. Data distribution (Alelyani et al., 2011) 8. Features correlation (Bhatt et al., 2012; Oreski et al., 2017) 9. Multivariate normality (Oreski et al., 2017) 10. Mean Kurtosis of attributes (Bhatt et al., 2012) 11. Mean skewness of attributes (Bhatt et al., 2012)
Data quality criteria	12. Accuracy (Omri et al., 2021; Redman, 1997) 13. Completeness (Omri et al., 2021; Oreski et al., 2017) 14. Consistency (Redman, 1997) 15. Timeliness (Redman, 1997)
Application related criteria	16. Field of application 17. Data source

For example, the data distribution a negligible impact on the data acquisition process. However, this characteristic is very important in the data processing and exploitation processes (diagnostic and prognostic). Thus, the importance of each of the criteria will be judged with respect to every PHM process separately.

3.3. Criteria / Sub-criteria weights determination

Notation:

- P_i : Elementary process i (i=1,..., L)
- D_h : Similar dataset h (h=1,..., Q)
- Cj : Criterion j (j=1,..., N)
- $X_{j,i}$: Weight of criterion j for process i
- SCk : Sub-criterion k (k=1,..., M)

- $Y_{k,i}$: Weight of sub-criterion k for process i
- W_k : Weight of sub-criterion k
- M_j : Number of sub-criterion related to the criterion j
- $Z_{h,k}$: Similarity score of the candidate dataset h with the target dataset with respect to the sub-criterion k
- R_h : Similarity score of the candidate dataset h with the target dataset.

In the AHP technique, a ratio scale is used to derive, two by two, the criteria's and sub-criteria's importance. This comparison, unlike techniques that use interval scales, requires no units (Ishizaka & Labib, 2011) and assures a more accurate decision than comparing all the criteria at once.
The pairwise comparison of criteria, and every group of sub-criteria, is realized using Saaty's 1-9 scale for pairwise comparison (Saaty, 2005) described in Table 2.

Table 2. Saaty's 1-9 scale for pairwise comparison

Intensity of importance	Definition
1	Equal importance
3	Moderate importance
5	Strong importance
7	Very strong importance
9	Extreme importance

To determine the weights of N criteria for the elemental process Pi, An NxN matrix is created, where $a_{j1,j2}$ describes the importance of criterion $Cj1$ over criterion $Cj2$. Therefore, for all j1 and j2, $a_{j1,j2}$ is the inverse of $a_{j2,j1}$ and $a_{j1,j1} = 1$.

$$\begin{bmatrix} 1 & a_{1,2} & .. & a_{1,N} \\ a_{2,1} & 1 & .. & a_{2,N} \\ .. & .. & .. & .. \\ a_{N,1} & .. & a_{N,N-1} & 1 \end{bmatrix} \quad (1)$$

This procedure is carried out to deduce the relative importance of the criteria by comparing them two by two using the fundamental scale of importance intensity. The weight $X_{j,i}$ of criterion Cj in relation to the process Pi is calculated using equation 2.

$$X_{j,i} = \frac{\sum_{j2=1}^{N} \left(\frac{a_{j1,j2}}{\sum_{j1=1}^{N} a_{j1,j2}}\right)}{N} \quad (2)$$

Similarly, the sub-criteria relating to each criterion are compared two by two, and the weight of each sub-criterion in relation to the Pi process is calculated using equation 4

$$\begin{bmatrix} 1 & b_{1,2} & .. & b_{1,Mj} \\ b_{2,1} & 1 & .. & b_{2,Mj} \\ .. & .. & .. & .. \\ b_{Mj,1} & .. & b_{Mj,Mj-1} & 1 \end{bmatrix} \quad (3)$$

$$Y_{k,i} = \frac{\sum_{k2=1}^{Mj} \left(\frac{b_{k1,k2}}{\sum_{k1=1}^{Mj} b_{k1,k2}}\right)}{Mj} \times X_{j,i} \quad (4)$$

At the end of this procedure, the weight of each criterion/sub-criterion is given, showing their importance for each elementary process of the PHM cycle.
In order to deduce the weight of a sub-criterion for the whole cycle, an average of these weights is calculated (equation 5).

$$W_k = \frac{\sum_{i=1}^{L} Y_{k,i}}{L} \quad (5)$$

3.4. Similarity score calculation

In this final step, the similarity $Z_{h,k}$ of every candidate dataset D_h with the studied dataset regarding each sub-criterion k is evaluated. The assessment is done using the fundamental scale of importance intensity (Table 2).
For quantitative criteria, an odd number between 1 and 9 is assigned, depending on the decision-maker's expertise. On the other hand, for qualitative criteria, only two possible values can be given, 9 for two data sets with similar attributes and 1 otherwise.
Finally, a normalized similarity score of each candidate dataset R_h is calculated using equation 6. The higher the similarity score, the more the concerned dataset is similar to the target dataset. A similarity score of 1 means that the two compared datasets have identical characteristics.

$$R_h = \frac{\sum_{k=1}^{M} Z_{h,k} \times W_k}{9} \quad (6)$$

4. ILLUSTRATIVE CASE STUDY

The proposed database similarity assessment methodology will be applied to a case study of bearing failure databases available online.
A bearing is a machine component that lessens friction between moving elements in mechanical engineering. It is frequently used in wheels or axles to support and guide a rotating or oscillating shaft. Bearings can be subject to various failures, manifested by cracks, wear marks, chips, and abnormal noises. These failures can significantly affect the mechanical and energy sectors' capacity to operate, level of safety, and financial aspect (Nectoux et al., 2012).
In the context of PHM applications for bearing condition prognosis, (Nectoux et al., 2012) provided a database for the IEEE PHM 2012 Prognostic Challenge. The experiments were car-

ried out on the PRONOSTIA platform at the Femto-ST Institute, and the results present 9 features relating to run-to-failure tests of 17 bearings.

4.1. Proposed criteria and PHM cycle modeling

For the application under consideration, based on the criteria summary table (Table 1), eleven sub-criteria split among three criteria were proposed. The sub-criteria relating to the standard criteria were retained, except for the number of classes. This selection is justified by the studied databases, which were not originally designed for classification purposes and lack class labels. In addition, the application-related criteria were also retained with the proposal of two additional criteria specific to this application, namely the number of operating conditions applied and the number of tested bearings. Moreover, two of the data quality sub-criteria were used in this case study. The completeness was evaluated as the ratio of non-empty cells over all available cells, and the accuracy was assessed as the presence or absence of noise.
For this application, the PHM cycle was simplified to 3 elementary processes, namely the data acquisition, the data preprocessing, and the prognostics processes.

4.2. Similar databases collection

A collection of four databases, available online, for the same applications, was carried out.
The first dataset (Kaggle, 2023) is provided by Quantum company in collaboration with Kharkiv Polytechnic Institute. It consists of 3-axis vibration measurements of 112 rotating bearings.
The second dataset (Qiu, Lee, Lin, & Yu, 2006) is a run-to-failure dataset of four bearings under one operating condition, provided by Qiu et al. Eight features related to the vibration and the temperature of the bearings were collected to study their health state.
The third data set (CWRU, .) is provided by Case Western Reserve University and presents ten statistical features related to measurements of 21 bearings under fixed operating conditions that manifested ten possible types of faults.
Finally, the fourth database presents recordings of the acceleration of a high-speed bearing used for wind turbines over 30 days (6 seconds daily). These recordings were made under two operating conditions.
Table 3 details the selected dataset characteristics in relation to the criteria and sub-criteria chosen for the study.

4.3. Criteria and sub-criteria weights calculation

As mentioned in section 3, and in order to determine the sub-criteria weights, a pairwise comparison of the importance of the criteria for every elementary PHM process was performed.
Table 4 details the process of comparing importance and calculating criteria weights for the data acquisition process.
The criteria weights were calculated using equation 2, after constructing the comparison matrix. The application criterion contributes the most to selecting a similar dataset for the data acquisition process. In addition, a similarity in the application criterion is strongly preferred to the quality criterion. In fact, a different application may require another data acquisition system. Moreover, as seen in Table 4, no two criteria are of equal importance for the acquisition process.
Table 5 compares the criteria importance and weights in the data preprocessing process. In contrast to the data acquisition process, the application criterion has an insignificant weight compared to standard and quality criteria, indicating a lower priority in this context. Conversely, the quality criterion holds the highest significance in selecting an appropriate dataset in the preprocessing process, holding nine times more importance than the application criterion and three times more significance than the standard criterion. These findings align with expectations, as the preprocessing process rarely depends on applications and focuses mainly on data quality and standard characteristics.
The weights of each family of sub-criteria were then determined for each elementary process of the PHM cycle. This is done by comparing them two by two using the 1-9 scale for pairwise comparison and then, by applying equation 4 to incorporate the weights of the associated criteria.
Table 6 shows the weights of the standard sub-criteria for the data acquisition process. The number of features is found to be the most important sub-criterion to assess the similarity between two datasets concerning the data acquisition process. In fact, features (variables) are collected using acquisition devices like sensors. These devices are costly and require studies to set them up and to ensure data acquisition. This sub-criterion is therefore the most important for this PHM process. The number of features sub-criterion is considered to be very strongly important than the number of instances, extremely important than the data extension sub-criterion, and moderately important than the data format sub-criterion.
The data format is the second most important standard sub-criterion to assess datasets similarity in relation to the data acquisition process. It is strongly more important than the number of instances and the number of binary features, and moderately more important than the data extension sub-criterion.
The Standard sub-criteria importance and weights for the prognostic process are described in Table 7. Similarly to the acquisition process, the number of features criterion is the most important factor in determining databases' similarity in relation to the prognostic process. Additionally, the data extension is the least impacting factor in both processes. The second most important standard sub-criterion is the number of binary features in this context. It is moderately more important than the number of instances and data format criterion and highly more important than the data extension criterion.
The final weights of all the considered sub-criteria for the

Table 3. Collected datasets characteristics

	Target dataset	**Candidate Dataset 1**	**Candidate Dataset 2**	**Candidate Dataset 3**	**Candidate Dataset 4**
Number of attributes	7	13	8	10	2
Number of instances	18196480	10265700	4415488	2048	29,286,800
Number of binary features	0	0	0	0	0
Data format	tabular	tabular	text	tabular	binary data container
Data extension	.csv	.csv	text	.csv	.mat
Field of application	Academic	Industrial	Academic	Industrial	energy industry
Data source	test bench	test bench	test bench	test bench	real life
Number of operating conditions	3	3	1	1	2
Number of bearings tested	17	112	4	21	1
Completeness	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %
Accuracy	noised	X	noised	X	noised

Table 4. Criteria matrix and weights for the data acquisition process

	Standard	Application	Quality	**Criteria weights**
Standard	1	1/3	3	**0,2605**
Application	3	1	5	**0,6333**
Quality	1/3	1/5	1	**0,1062**

Table 5. Criteria matrix and weights for the preprocessing process

	Standard	Application	Quality	**Criteria weights**
Standard	1	7	1/3	**0,2946**
Application	1/7	1	1/9	**0,0567**
Quality	3	9	1	**0,6486**

whole PHM cycle are detailed in Table 8.

The accuracy sub-criterion is found to be the most important. Since the scores are normalized, then a weight of 0,3603 means that this sub-criterion contributes by 36,03% to the final decision about datasets similarity. The following sub-criteria are the number of tested bearings, the number of operating conditions, and the number of collected features. These four sub-criteria contribute by more than 75% to the final decision.

4.4. Similarity score calculation and decision-making

In this final step, the similarity of every candidate dataset with the target dataset is evaluated with respect to every sub-criterion using the fundamental scale of importance intensity (Table 2).

Similarity based on qualitative criteria is assessed using the Saaty scale. In other words, if both datasets have the same attribute, a rating of 9 is assigned; otherwise, a rating of 1 is assigned.

For example, for the data format sub-criterion, two candidate datasets are of the same format as the target dataset, so they got a similarity score of 9. The other two datasets are of different formats (text and binary data container), leading to a weak similarity score of 1.

A score of similarity, according to every sub-criterion, between each candidate dataset and the target dataset is given. Afterward, the weights deducted in the previous step are used to get a similarity score for every candidate dataset 8.

The second dataset (Qiu et al., 2006) is found to be the most similar dataset to the target dataset (Nectoux et al., 2012) with a similarity score of 0,7143. However, the first candidate dataset (Kaggle, 2023) is the least similar dataset to the target dataset. This is mainly caused by the difference in the accuracy sub-criterion, the number of tested bearings and the number of features. These sub-criterion were found to be three of the four most important comparison sub-criteria. Therefore, a low score in these attributes leads to a weak overall similarity score.

If a simple normalized mean of the similarity scores is calculated, the first candidate dataset will obtain a higher score of 0,7172, meaning that it is the most similar dataset. This shows the importance of assigning weights to the comparison sub-criteria.

It is important to note that, although the fourth dataset is the only one originating from the real world, it was not selected. This decision stems from the fact that the 'data source' criterion is just one of several simulation criteria used in the selection process. Moreover, the target dataset itself is derived from a simulation, rendering dataset number 4 dissimilar in terms of data source. Our aim is to select the dataset that most closely resembles the target dataset, rather than simply identifying the best dataset.

5. DECISION RELIABILITY

The methodology outlined in this study hinges upon conducting pairwise comparisons of both criteria and sub-criteria to ascertain their respective weights. These comparisons are based on subjective judgments provided by the decision-maker. Consequently, it becomes important to assess the consistency of these judgments. Consistency, within the context of the

Table 6. Standard sub-criteria matrix and weights for the data acquisition process

	Number of instances	Number of features	Number of binary features	Data format	Data extension	**Criteria weights**
Number of instances	1	1/7	1/3	1/5	3	**0,0740**
Number of features	7	1	5	3	9	**0,5048**
Number of binary features	3	1/5	1	1/5	3	**0,1163**
Data format	5	1/3	5	1	3	**0,2581**
Data extension	1/3	1/9	1/3	1/3	1	**0,0468**

Table 7. Standard sub-criteria matrix and weights for the prognostic process

	Number of instances	Number of features	Number of binary features	Data format	Data extension	**Criteria weights**
Number of instances	1	1/5	1/3	3	7	**0,1449**
Number of features	5	1	3	7	9	**0,4992**
Number of binary features	3	1/3	1	3	7	**0,2298**
Data format	1/3	1/7	1/3	1	7	**0,0962**
Data extension	1/7	1/9	1/7	1/7	1	**0,0299**

Table 8. Final weights and similarly scores of the candidate datasets

Sub-criteria	Sub-criteria weights	**Candidate dataset 1**	**Candidate dataset 2**	**Candidate dataset 3**	**Candidate dataset 4**
Number of instances	**0,0156**	5	3	1	5
Number of features	**0,1008**	3	7	5	1
Number of binary features	**0,0224**	9	9	9	9
Data format	**0,0584**	9	1	9	1
Data extension	**0,0177**	9	5	9	1
Accuracy	**0,3603**	1	9	1	9
Completeness	**0,0721**	9	9	9	9
Field of application	**0,0354**	5	9	5	3
Data source	**0,0276**	9	9	9	5
Number of operating conditions	**0,1229**	9	3	3	7
Number of bearings	**0,1668**	3	3	7	1
Similarity score		**0,4787**	**0,7143**	**0,4863**	**0,6244**

Analytic Hierarchy Process (AHP), denotes the extent to which the pairwise comparisons rendered by decision-makers exhibit logical coherence and absence of contradictions. Inconsistencies in judgments bear the risk of yielding unreliable weight assignments, thereby potentially skewing the subsequent similarity evaluations.

Several works have addressed the problem of consistency of AHP matrices. One way to deal with this is by determining the Consistency Ratio (CR) (Pant, Kumar, Ram, Klochkov, & Sharma, 2022; Franek & Kresta, 2014). First, the Consistency Index (CI) is computed according to equation 7:

$$CI = \frac{\lambda_{max} - N}{N - 1} \tag{7}$$

with λ_{max} representing the largest eigenvalue of the pairwise comparison matrix and N indicating the matrix size (number of criteria or sub-criteria).

Using pre-defined tables (Table 9), the Random Index (RI) corresponding to the matrix size is determined. The Consistency Ratio (CR) is then calculated by dividing CI by RI. A CR value below 0.1 signifies acceptable consistency in judgments, while values exceeding 0,1 may indicate potential inconsistencies requiring further scrutiny or adjustment.

As an example, the consistency ratio of the criteria pairwise comparison matrix is 0,03 for the data acquisition process and 0,07 for the preprocessing process. These results demonstrate that the weights of the resulting criteria are consistent and can be used to reliably determine criteria weights.

On the other hand, the consistency ratio of the standard sub-criteria matrix for the prognostic process is 0,11 meaning that the comparison need to be adjusted in order to get a consistent judgment of the sub-criteria weights.

6. CONCLUSION

In this work, a database comparison approach was proposed to find a solution to the problem of lack of data for PHM applications. Indeed, for this field of study, and in order to develop a data-driven PHM solution, datasets need to be available, containing all the variables describing the system under study and complying with quality standards. In reality, this is

Table 9. Random index for the AHP consistency ratio (Saaty, 1980)

Number of rows	1	2	3	4	5	6	7	8	9
RI	0	0	0,58	0,90	1,12	1,24	1,32	1,41	1,45

not always the case.

Therefore, we have proposed an approach for assessing the similarity between a target dataset and a set of datasets available online. A set of criteria has been proposed, based on data characteristics. As the criteria are not equally important for judging similarity, a weight for each criterion is determined using the analytical hierarchy process. The similarity of the datasets is then scored against each criterion, and a normalized score is calculated for each dataset.

The proposed approach has been applied to an illustrative case, where the similarity of four datasets with a bearing operating database has been evaluated. The application leads to calculating similarity scores for each dataset and selecting the most similar one.

This work presents a first step towards solving the problem of lack of data for PHM applications. It makes it possible to design a PHM solution for a given system without the need to use data directly from that system.

On the other hand, this proposal is limited by the subjectivity of the decision-maker. The latter, responsible for rating similarity and judging the importance of criteria, may be biased and lead to subjective decisions. We therefore recommend that weights and scores be allocated by several experts at the same time, in order to limit the subjectivity of decision-makers.

In addition, considering the limitations of our current methodology, future studies may employ fuzzy techniques to reduce the uncertainty of the decision. Furthermore, in this work, we proposed a non-exhaustive list of criteria, other criteria can also be used, namely the maturity of the data for example, which leads to the generalization of the approach to various fields and applications.

REFERENCES

Ahmadi, A., Arasteh Khouy, I., Kumar, U., & Schunnesson, H. (2009). Selection of maintenance strategy, using analytical hierarchy process. *Communications in Dependability and Quality Management*, *12*(1), 121–132.

Alelyani, S., Liu, H., & Wang, L. (2011). The effect of the characteristics of the dataset on the selection stability. In *2011 ieee 23rd international conference on tools with artificial intelligence* (pp. 970–977).

Benjelloun, O., Chen, S., & Noy, N. (2020). Google dataset search by the numbers. In *International semantic web conference* (pp. 667–682).

Bhatt, N., Thakkar, A., & Ganatra, A. (2012). A survey & current research challenges in meta learning approaches based on dataset characteristics. *International Journal of soft computing and Engineering*, *2*(10), 234–247.

Bougacha, O., Varnier, C., & Zerhouni, N. (2022). Impact of decision horizon on post-prognostics maintenance and missions scheduling: a railways case study. *International Journal of Rail Transportation*, *10*(4), 516–546.

Brunelli, M. (2014). *Introduction to the analytic hierarchy process*. Springer.

Cabrita, M. D. R., & Frade, R. (2016). Supplier selection approach: integrating analytic hierarchy process and supplier risk analysis. *International Journal of Business and Systems Research*, *10*(2-4), 238–261.

CWRU. (.). *Case western reserve university bearing data center dataset.* `https://engineering.case.edu/bearingdatacenter`. (Accessed on March 12, 2024)

Franek, J., & Kresta, A. (2014). Judgment scales and consistency measure in ahp. *Procedia economics and finance*, *12*, 164–173.

Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2018). Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics*, *66*(9), 7316–7325.

Ishizaka, A. (2004). The advantages of clusters in ahp. In *The 15th mini-euro conference, mudsm.*

Ishizaka, A., & Labib, A. (2011). Review of the main developments in the analytic hierarchy process. *Expert systems with applications*, *38*(11), 14336–14345.

Kaggle. (2023). *Bearing classification dataset.* `www.kaggle.com/datasets/isaienkov/bearing-classification`. (Accessed on March 12, 2024)

Kilic, H. S., Zaim, S., & Delen, D. (2014). Development of a hybrid methodology for erp system selection: The case of turkish airlines. *Decision Support Systems*, *66*, 82–92.

Li, X., Zhang, W., Ding, Q., & Sun, J.-Q. (2020). Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. *Journal of Intelligent Manufacturing*, *31*(2), 433–452.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. In *Ieee international conference on prognostics and health management, phm'12.* (pp. 1–8).

Nydick, R. L., & Hill, R. P. (1992). Using the analytic hierarchy process to structure the supplier selection procedure. *International Journal of purchasing and materials management*, *28*(2), 31–36.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., & Zerhouni, N. (2020). Industrial data management strategy towards an sme-oriented phm. *Journal of Manufacturing Systems*, *56*, 23–36.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., & Zerhouni, N. (2021). Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in industry*, *127*, 103414.

Oreski, D., Oreski, S., & Klicek, B. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, *52*, 109–119.

Pant, S., Kumar, A., Ram, M., Klochkov, Y., & Sharma, H. K. (2022). Consistency indices in analytic hierarchy process: a review. *Mathematics*, *10*(8), 1206.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Qiu, H., Lee, J., Lin, J., & Yu, G. (2006). Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of sound and vibration*, *289*(4-5), 1066–1090.

Redman, T. C. (1997). *Data quality for the information age*. Artech House, Inc.

Ren, J., & Lützen, M. (2015). Fuzzy multi-criteria decision-making method for technology selection for emissions reduction from shipping under uncertainties. *Transportation Research Part D: Transport and Environment*, *40*, 43–60.

Saaty, T. L. (1980). The analytic hierarchy process (ahp). *The Journal of the Operational Research Society*, *41*(11), 1073–1076.

Saaty, T. L. (2005). *Theory and applications of the analytic network process: decision making with benefits, opportunities, costs, and risks*. RWS publications.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).

Shao, S., McAleer, S., Yan, R., & Baldi, P. (2018). Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, *15*(4), 2446–2455.

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical systems and signal processing*, *64*, 100–131.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Tobon-Mejia, D. A., Medjaher, K., Zerhouni, N., & Tripot, G. (2012). A data-driven failure prognostics method based on mixture of gaussians hidden markov models. *IEEE Transactions on reliability*, *61*(2), 491–503.

Wang, Z., Yang, J., Jiang, H., & Fan, X. (2020). Cnn training with twenty samples for crack detection via data augmentation. *Sensors*, *20*(17), 4849.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1–40.

Wen, L., Gao, L., & Li, X. (2017). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on systems, man, and cybernetics: systems*, *49*(1), 136–144.

A PHM implementation framework for MASS (Maritime Autonomous Surface Ships) based on RAM (Reliability, Availability, Maintainability) analysis

Toby Adam Michael Russell[1], Octavian Niculita[2]

[1]*Ocean Infinity Ltd, UK*

toby.russell@oceaninfinity.com

[1,2]*Glasgow Caledonian University, Glasgow, Scotland*
octavian.niculita@gcu.ac.uk

ABSTRACT

The paper focuses on PHM in the maritime industry, specifically on the maintenance of uncrewed vessels, in contrast to the more commonly discussed navigation. The paper examines the potential challenges of removing the maintenance crew and the potential benefits that can result from this major change in operations.
The removal of the primary maintenance team from a vessel necessitates an increase in monitoring and analysis that can be realised by the techniques of PHM. By looking from the perspective of stakeholders, the challenges and opportunities of PHM implementation become clearer. In comparing the challenges that faced other industries with the maritime industry, roadmaps and proposals can be drawn up for vessel owners. There is a correlation between the phased removal of the engineering crew and the increases in monitoring that is required. Current large vessels that do not carry passengers can operate with UMS (un-manned machinery space) for limited periods. To allow this a specific set of sensors referred to as E0 (Engineers-zero) must be established and maintained. This E0 sensor set forms the basis for what is needed to allow UMS for longer periods of time. The critical equipment, as deemed by class societies, is monitored by E0. Acquiring the data from the E0 sensor set and performing PHM analysis on the data allows remote engineers to accurately determine the current and future state of critical equipment. This equipment list needs to be expanded. Causality based risk modelling is employed to establish a data driven critical equipment list and minimum sensor set to cover the maximum amount of failure modes. This builds on the current required E0 sensor set.
With a conventional maintenance system onboard a vessel the crew are doing a lot of the sensing. The crew act as intermediaries between various systems, taking data from one system to help diagnose another system, making a change to one system to help improve another system. The maintenance crew must balance the interfaces of each system so that a harmony or equilibrium can be achieved. This balancing act is part of what makes a PHM study on a vessel so interesting. Many systems onboard a vessel have a sole purpose to support the crew. With the removal of the crew these support systems can also be removed, simplifying the overall engineering of the vessel.
The methodology that has been used to assess the above points is to create a framework for the design and deployment of PHM to marine assets. The framework relates to RAM (Reliability, Availability, Maintainability) and considers stakeholder points of view and their inputs' implications. In developing the framework, the stakeholder group is realised. The framework compares the 'As-Is' conventional method against the proposed PHM framework. The conclusions are that the E0 philosophy can be expanded upon to facilitate the integration of PHM. Also, the paper concludes that a PHM deployment framework gives the maritime industry a basis for using this modern technique for machinery health. Lastly, the paper shows that PHM is a vital element to uncrewed vessels.

1. INTRODUCTION

In this paper we will investigate a way to use the principles of RAM (Reliability Availability Maintainability) to facilitate uncrewed vessel operations. The focus of the paper is going to be only on the uncrewed Engineering operations, not on the Navigational operations.

The motivations for conducting this study are,

1. The maritime industries move to leverage modern technology to reduce environmental impacts of shipping is a major motivator for this study.
2. To offer a method for enabling remote engineering of uncrewed vessels both in terms of the RAM driven PHM system and organisational architecture.
3. To increase safety at sea by reducing the number of crew needed to locally operate the vessels.

To validate the decision to move to a RAM enabled PHM maintenance system, it is important to carry out a comparison between different maintenance systems. To do this, key performance indicators (KPI) need to be established to track and compare the differences in performance of each, also, standardizing resources needed into categories.

Table 1: Generic Maintenance System KPI's

KPI	Description
Availability	Actual and predicted availability are critical to commercial operations
Human reliance / human error	How reliant the system is on individual humans
Unplanned maintenance tasks	Quantity and frequency of unplanned maintenance tasks
Planned maintenance tasks	Quantity and frequency of planned maintenance tasks
Set up cost	Cost of setting up the maintenance system
Running costs	Cost of running the maintenance system
Maintenance Costs	Cost of maintaining the asset during its lifetime

Table 1 above shows a list of KPI's that can be used to compare maintenance systems. This is not a definitive list, but serves as a overarching view of a maintenance systems efficiency.
Each of the evolutions of the proposed maintenance systems will be compared on the above KPI's with results in the appendix of this paper.
Other metrics that can be considered are:

- Time
- Environmental
- Complexity
- Sustainability

The maintenance landscape in the maritime industries is due for an overhaul. Remote Engineering is becoming a factor. This paper aims to show a possible way forward

This paper uses a typical 70 – 100 meter length overall vessel designed to have an operational lifespan of 20 years as an example asset.

To recap on the evolution of maintenance so far, from 1920 to now we can see that cost and availability have always been the biggest drivers. As a high level over view the following is a view of these evolutions. Figure 1 shows how maintenance systems have progressed since "run to failure" where a machine was only repaired, rather than maintained. To Dynamic Maintenance, where failures are identified in a juvenile state, allowing operators to plan corrective action and impose mitigating actions to prolong the asset life prior to maintenance.

2. MAINTAINING A VESSEL

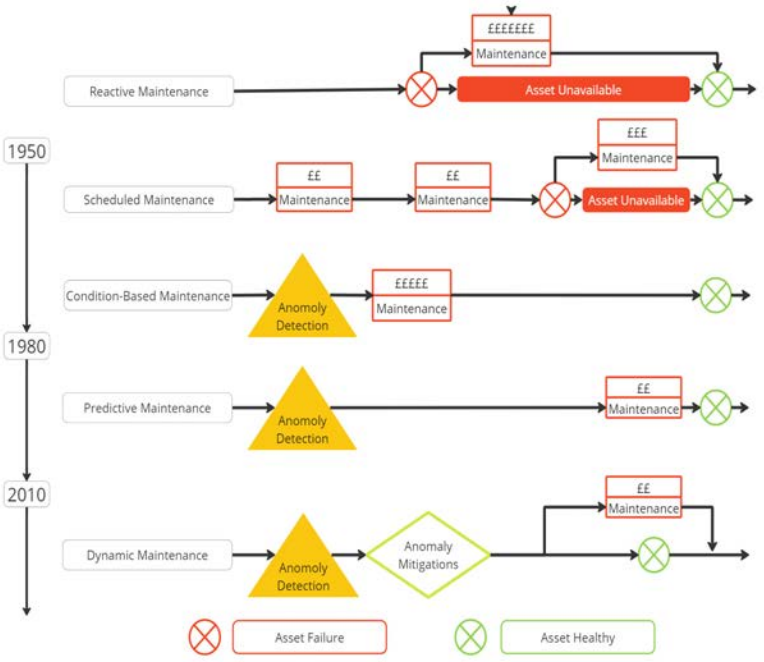


Figure 1. Evolution of Maintenance Systems

Figure 1 also shows that as the system evolves maintenance can be done at targeted times when it is most effective, this optimises the maintenance and improves asset availability. This optimisation culminates in a system that can mitigate the failure to prolong asset life so that maintenance intervals can be fixed.

Classification societies guide and ensure that ships are maintained to certain standards. For ships, classification societies provide classification services that involve assessing the structural integrity, safety, and performance of

vessels according to their rules and regulations. These services include:

- surveys at various stages of a ship's life, including during construction, after delivery, and during operation. these surveys verify that the ship complies with class rules and standards.
- regular surveys are conducted to ensure that the ship remains in compliance with class requirements throughout its operational life. these surveys cover aspects such as hull integrity, machinery, electrical systems, and safety equipment.
- guidelines and requirements for implementing maintenance management systems onboard ships are also provided by classification societies. these systems help shipowners and operators manage maintenance activities effectively to ensure the continued safe operation of the vessel.
- condition monitoring and predictive maintenance, allowing ship operators to identify potential issues before they lead to failures or downtime.
- assistance and support to shipowners in understanding and complying with relevant regulations and standards, including international conventions and flag state requirements.

2.1. Classification Society Methodology

Classification societies assist ship owners in improving maintenance to support increased safety. This starts by assigning critical equipment. Criticality assessment is carried out to comply with standards such as NORSOK Z-008 PSA, ISM code 10.3 and OVMSA. For the analysis, generally, all onboard maintainable items are included. A risk-based assessment is then carried out in terms of impact to health, environment, operation, property, MASS capability.

PHM is currently included in Class Society documentation, but there is little in the way of guidance. The generic view is that PHM can provide valuable information for corrective and preventative action, inclining operations adjustment (Shipping, 2018).

Paper based analysis is used (rather than model based) to risk assess equipment and deem it critical if needed. While RAM is not used in its entirety by classification societies, they are considering reliability and safety analysis as part of a certification process of a vessel.

Classification Societies also issue vessels with ‘notations’ which indicate to clients the level of quality or performance the vessel has achieved. DNV for example issue a Condition Monitoring System notation if the vessel can prove compliance to specific equipment health requirements. At time of writing there is no PHM notation.

3. UNMANNED MACHINERY SPACES (UMS)

Unmanned Machinery Spaces (UMS) are engine rooms or machinery spaces on ships that are designed and equipped to operate without the continuous presence of personnel. This means that machinery and systems are automated and monitored remotely from a control room, reducing the need for crew members to be physically present in these spaces. One of the classifications and notations associated with UMS is the E0 (Engineers Zero) notation.

3.1. Key Features of UMS with E0 Notation:

Advanced Automation: UMS with E0 notation feature advanced automation systems that control and monitor machinery and systems in the engine room. Redundant systems and fail-safe mechanisms are implemented to ensure continuous operation and minimize the risk of failure.

Machinery and systems are monitored remotely from control rooms or other locations onboard the ship. Automated alarm and alert systems notify onboard personnel or shore-based monitoring centres of any abnormalities or emergencies.

Emergency procedures and backup systems are in place to respond to emergencies or system failures, including the ability to remotely intervene or override automated systems if necessary.

Ships with UMS, including those with the E0 notation, must comply with relevant regulations and guidelines governing unmanned or partially unmanned operations, such as those issued by the International Maritime Organization (IMO) and flag state authorities.

4. ENGINEERS ZERO (E0)

For a typical vessel between 70 and 100 meters, the E0 monitoring and alarm list is a comprehensive set of internal alarms for various systems and components on a ship. These alarms are designed to monitor the status of critical machinery, systems, and equipment, and they provide alerts in case of abnormalities or failures. The following is a breakdown of the categories and some examples of alarms existent on UMS with E0 capability.

Internal Alarms: These alarms are related to the ship's internal systems and components.

Earth Failure Alarms: These alarms indicate a potential earth (ground) failure in specific components, such as controllers and power supplies.

Power Failure Alarms: These alarms notify of power failures in specific components, such as controllers and power supplies.

Fuel Oil System Alarms: These alarms monitor the fuel oil system, including tank levels, overflows, and pressure.

Main Propulsion Alarms (Port and Starboard): These alarms relate to the main propulsion systems, including power

supply failures, low oil pressure, high temperatures, and warnings regarding control systems.

Generator Set Alarms: These alarms monitor various parameters of generator sets, including fuel levels, water pressure, oil pressure, temperatures, overspeed, and abnormal conditions.

Lube Oil System Alarms: These alarms monitor the lube oil system, including separator alarms and overflow alarms.

Cooling System Alarms (Sea Water and Fresh Water): These alarms monitor the cooling systems, including low pressure alarms for sea water and freshwater systems and low-level alarms for expansion tanks.

Compressed Air System Alarms: These alarms monitor the compressed air system, including low-pressure alarms for starting air receivers and quick-closing cabinets.

Bilge System Alarms: These alarms monitor bilge levels in various compartments throughout the ship.

Main Switchboard Alarms: These alarms monitor the main switchboard for various failures and abnormalities in power supply and distribution.

Miscellaneous Alarms: These alarms cover a range of miscellaneous systems and components, including communication errors, black-out bus failures, and controller failures.

These alarms are crucial for maintaining the safe and efficient operation of the vessel by promptly alerting crew members to any issues that may arise within the ship's systems. Typical maintenance intervals are having a monthly occurrence for E0 alarms.

Overall, there will be approximately 450 alarms across the systems described above. With all these in place, a vessel can apply for E0 notation and operate with unmanned machinery spaces for certain periods of time. The rules around when a vessel can operate in UMS are also not described in this paper.

5. EXPANSION ON E0 FOR UNCREWED SHIPS

Having the E0 notation in place allows the engineering staff to rely on the automation system to alert them to critical issues onboard. Expanding on this principle can form a basis for much longer periods of UMS operation. To build upon this principle to facilitate uncrewed operations the maintenance and monitoring strategy of ships needs to be changed.

5.1. Dynamic Positioning

It is also worth mentioning dynamic position (DP) systems in the context of this investigation. DP is used to keep a vessel in location. There are 3 grades for the DP capability, as shown below.

Table 2: Dynamic Positioning Grades

DP1	Position Keeping
DP2	DP1 + any single component failure and vessel remains 100% capable of position keeping
DP3	DP2 + Any single compartment failure and vessel remains 100% capable of position keeping

It is worth mentioning the DP grades here because they require a certain level of redundancy. To be classed as a DP3 vessel the vessel must be designed in such a way that if any compartment is lost the vessel is still 100% capable. This could be a fire in an engine room for example. The philosophies on DP grade can also be expanded and contribute to a foundation for RAM of uncrewed vessels.

6. VESSEL MAINTENANCE

Vessel maintenance covers the vessel itself and the systems installed onboard. There are systems necessary for the vessel to operate, and to support the crew. The SFI coding system is used in most cases to group the maintenance tasks into the following parent categories. Table 3 below outlines the high level SFI categories and the average number of tasks in each.

Table 3: - tasks per vessel system category

Category	Average Number of Tasks
1. Ship General	20-50
2. Hull	150-200
3. Equipment for Cargo	10-50
4. Ship Equipment	200-600
5. Equipment for Crew and Passengers	300-500
6. Machinery Main Components	100-500
7. Systems for Machinery Main Components	300-600
8. Ship Common System	100-300
9. Payload Equipment.	Depends on Payload types.

7. CONVENTIONAL MAINTENANCE SYSTEMS

Here we look at the typical modern conventional method of maintaining a vessel.

A conventional vessel uses a combination of planned, preventative, corrective and breakdown maintenance strategies. The two primary sources of information are the PMS (Planned Maintenance System) and the Engineers observations. A general overview of this process is shown in figure 2 below.

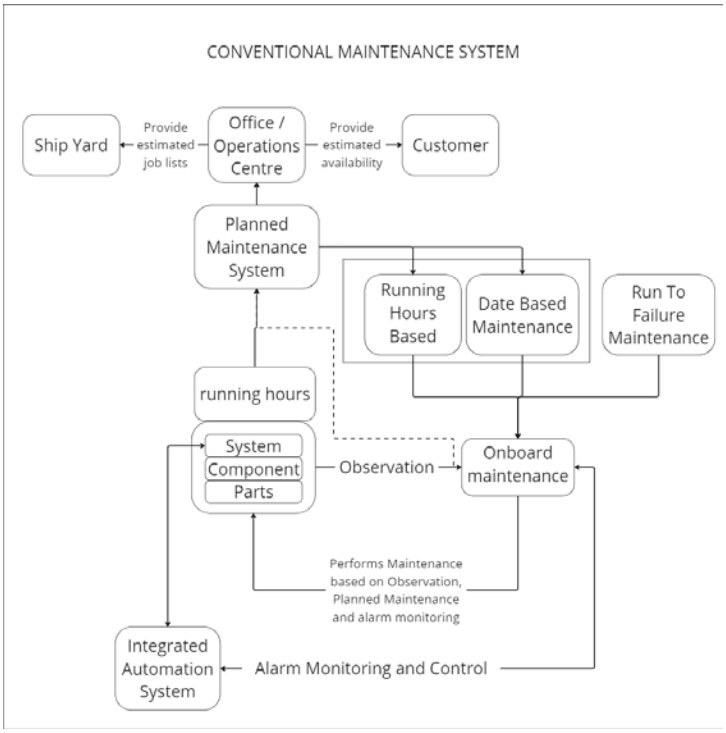


Figure 2. Conventional Maintenance System

7.1. Conventional Maintenance System Description

The system shown in figure 2 has the relationship between the onboard maintenance team and the asset to be maintained at the centre. Physical metrics of the asset are relayed to the engineers by an integrated automation system. This system also facilitates control of the asset. The assets running hours are maintained in the PMS which issues jobs to the onboard maintenance team. The team also react to faults and failures as they are observed. The PMS is linked to a shoreside system that assists the "office" in planning yard periods and vessel's availability to clients. The effectiveness of this system relies on the team onboard.

7.2. Method of Performing Maintenance

Maintenance is generally performed by an onboard team of engineers. If more extensive maintenance is needed specialists from OEM's are brought in. with this maintenance strategy there is a lot of reactive maintenance being done or planned maintenance that has no connection to the actual health of the asset.

Maintenance tasks are issued by the PMS and carried out by appropriate member of the engineering team. The onboard team monitor spares and consumable usage and submit orders when stock is running low. Certain items are classed as critical spares by a classification society to ensure that the safety critical equipment has spares onboard at all times.

Below is a summary of a typical maintenance team on a conventional vessel between 70 and 100 meters in length.

Table 4: Average Annual Cost of Engineering Team on Conventional Vessel

Position	Quantity	Qualification	Average annual cost
Motor Man	2	Marine Engineer operator license or similar	$70,000
Third Engineer	2	3 years college	$100,000
Second Engineer	2	4 years college	$150,000
Chief Engineer	2	5 years college	$210,000
Electro Technical Officer (ETO)	2	3 years college	$150,000

In Table 4. we can see that the total average cost of the onboard maintenance team, including flights and other travel is $680,000 over an anticipated life span of 20 years engineering crew costs amount to $13,600,000. This is one of the costs that uncrewed operations can mitigate.

A summary of the other resources required for a conventional maintenance system over a presumed 20 year life span.

Table 5: Example summary of average OPEX for conventional maintenance

Resource	Cost Over 20-year Asset Life Span
Engineering Crew Costs	$13,600,000
Dry Dock / docking	$2,000,000

Dry Dock specific Maintenance tasks	$6,000,000
OEM Maintenance	$10,000,000
General maintenance spares and consumables	$90,000,000
Total	$121,600,000

Table 5 shows average operational expenditure for maintenance of a typical lean crewed 70m-100m length overall vessel.

The yearly quantity of maintenance tasks varies through an assets lifetime, as can be seen in Figure 3 below.

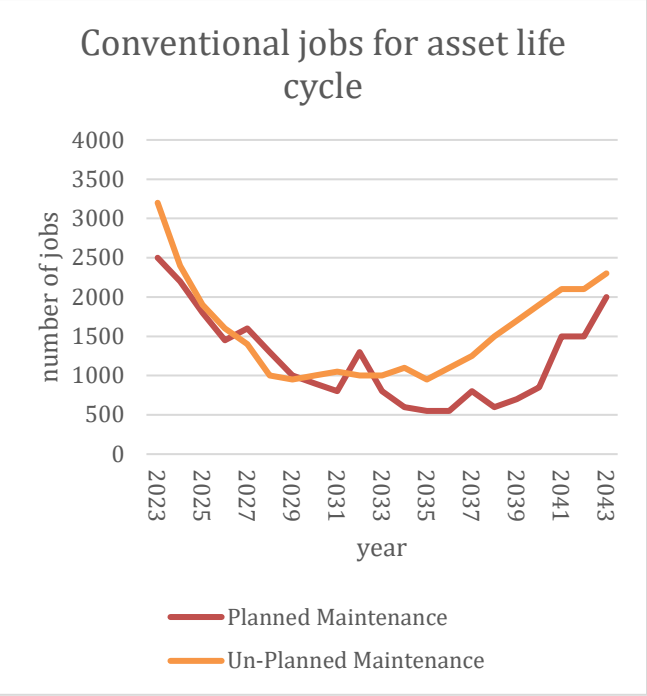


Figure 3. Number of maintenance tasks over asset lifespan

Figure 3 shows the average number of jobs for both planned and unplanned maintenance tasks. The scale is from first launch to end of life for the vessel, the vessel in the example is designed to have a 20 year life span. The general trend is a high number of jobs that decreases during the infancy of the asset, component failure during this period is often referred to as infant mortality. The trend flattens out after the first 2 to 3 years and then rises again as the asset starts to age. Figure 2 also shows some higher peaks in planned maintenance jobs, these are jobs that have been completed during the vessels 5 yearly dry docking, with this maintenance philosophy the 5 yearly dry docking is unavoidable down time, so grouping maintenance tasks to coincide with this improves efficiency.

With a conventional maintenance strategy as outlined in Figure 1 there is not much room to improve the efficiency of maintenance, there can be no reduction in spares, no extension of maintenance periods and no improvement in efficiency of systems for carbon footprint reduction. However, having a full complement of engineering staff onboard the asset means there is little need for maintenance management.

The data for number of planned jobs is taken from consolidation of OEM maintenance tasks that are recommended in the user manuals across all systems integrated to the vessel.

The data on unplanned maintenance tasks is obtained from historical entries in planned maintenance systems. Each task that is conducted onboard must be recorded in the planned maintenance system, whether it was a planned or unplanned task. Because of the nature of unplanned tasks the data is an average from historical data only.

7.3. Controls and Management

A conventional maintenance system is primarily managed by the chief engineer (CE) onboard the vessel. The CE keeps track of upcoming planned maintenance and ensures the correct spares will be available. The CE usually organises OEM maintenance also. There will be an office based team managing major planned maintenance intervals such as the 5 yearly dry docking, in cooperation with the onboard CE.

The controls that are in place to ensure maintenance is done correctly are mostly down to the CE onboard either checking the work or trusting the engineering team. Because the CE cannot check every single detail, they must be confident in the team. This is why the engineers onboard must have certificates of competency that are revalidated every 5 years, to ensure that they are still competent to perform the maintenance activities assigned to their role.

7.4. System Health Indicators and Key Performance Indicators

In a conventional maintenance system, the SHI are primarily observations by the crew. The Automation system can alert the crew to a parameter reaching a set point, for example if temperature increases to 50 degrees an alarm is triggered. The setpoints are controlled, especially for the E0 alarms. Once an alarm is triggered diagnosis is done by the maintenance crew using tools and human senses. Some trending is possible within the IAS on modern ships, although the majority of vessels operating may not be capable of trending a metric.

The SHI/KPI are selected following literature review and consolidating the metrics that are generally used. The values against each of the KPI/SHI are estimated based on achievements made by similar PHM implementation in other industries against historical data from vessel maintenance.

Table 6 below shows example KPI's / SHI's for a conventional maintenance system.

Table 6: Conventional Maintenance System Metrics

KPI / SHI	Description
Availability	292 days per year potential
Human reliance / human error	5 humans obtaining information for maintenance
Unplanned maintenance tasks	3200 tasks
Planned maintenance tasks	2500 tasks
Set up cost	$250,000 – one time cost
Running costs	$500,000 – yearly running costs
Maintenance Costs	$500, 000 – yearly spares / consumables

8. Conventional Condition Based or Reliability based maintenance System

The next step up from the maintenance system previously discussed starts to bring in condition monitoring. According to Lloyds Register only 17% of classed ships operate with an approved PMS, and only 12% of these use condition monitoring, leaving ~2% of classed ships with a condition monitoring system in place (Shorten, 2012). This means that as of 2012 only 12% of registered vessels operate using the Conventional Philosophy described in section 6.0 of this paper, and only 2% are using the system described in this section. This data is 14 years old at time of writing, so should be considered out of date, however this is the most recent formal data on the usage of CMS in the maritime industry.

Below is a generic example of including condition monitoring into the maintenance system of a vessel.

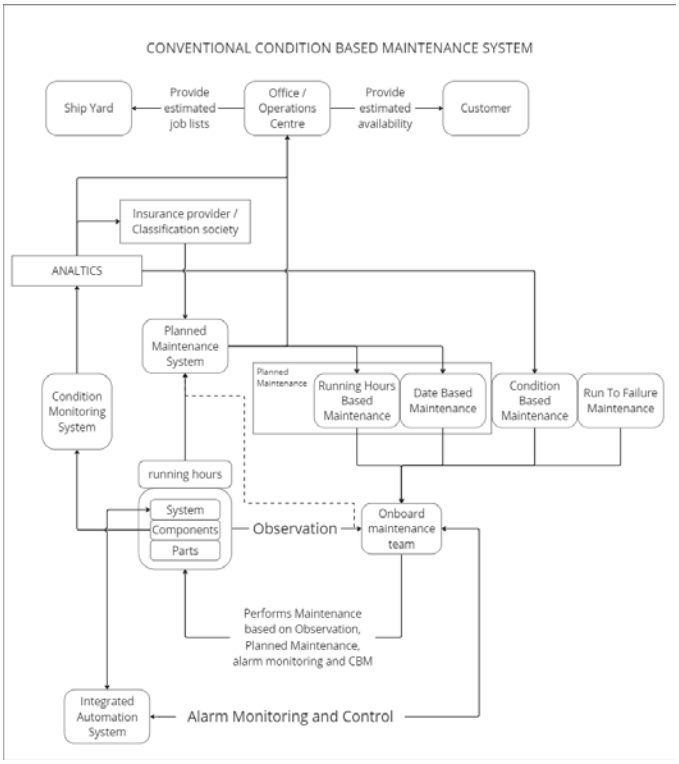


Figure 4. Conventional Condition Based Maintenance System layout

8.1. Conventional Condition Based Maintenance System Description

Building on the description in section 6.1, there is now the addition of a condition monitoring System (CMS) as can be seen in Figure 4. The output of the CMS is raw data that must be analysed. The "insights" that are generated from the analysis can be used to extend maintenance by sending to classification society and guiding the maintenance actions of the onboard team.

Correct application of CMS is vital for this maintenance system to work. Covering the asset with sensors is expensive and ineffective. The typical system of this type on vessels uses a vibration probe and measurements are taken at intervals. This method allows human error from the start as measurements are not always taken consistently with the asset in the same state. A RAM approach here delivers effective designs for CMS based on data driven reasons. Digital twins can be used for design of CMS for efficient sensor sets to cover the maximum amount of failure modes. Failure modes must be properly understood and categorized at this stage. Failure to properly design the CMS at this stage of evolution will increase costs and complexity of maintaining the asset. Typically, the CMS is applied to large rotating machines and vibration is the only sensing type.

8.2. Method of performing Maintenance.

Many CMS systems found in the maritime industry are not fixed, instead a portable sensor is used to take vibration readings. One of the maintenance crew is tasked with carrying out the measurement. This manual measurement introduces many human errors and means continuous monitoring is not possible. Once measurements are taken they are uploaded back to a land based office and then transferred to a 3rd party analytics provider. When potential faults are identified the company is informed who then inform the vessel and action can be taken. Due to the potential errors in measurement the results are often in accurate and so trust in the system does not develop. Often the system is abandoned and only the necessary readings are taken with little to no action from results.

8.3. System Health Indicators

As well as those mentioned in the previous iteration, this strategy can produce vibration health indicators.

Table 7: Conventional Condition Based Maintenance System Metrics

KPI / SHI	Description
Availability	292 days per year potential
Human reliance / human error	5 humans obtaining information for maintenance
Unplanned maintenance tasks	3000 tasks
Planned maintenance tasks	2800 tasks
Set up cost	$300,000 – one time cost
Running costs	$680,000 – yearly running costs
Maintenance Costs	$500, 000 – yearly spares / consumables

9. Prognostic Health Management maintenance System

In this section we look at the inclusion of prognostics in the maintenance strategy. It is important to note that this is where significant divergence occurs between a RAM enabled maintenance system and one that has not taken into account the nature of RAM.

Figure 5 below shows a general example layout of a PHMMS.

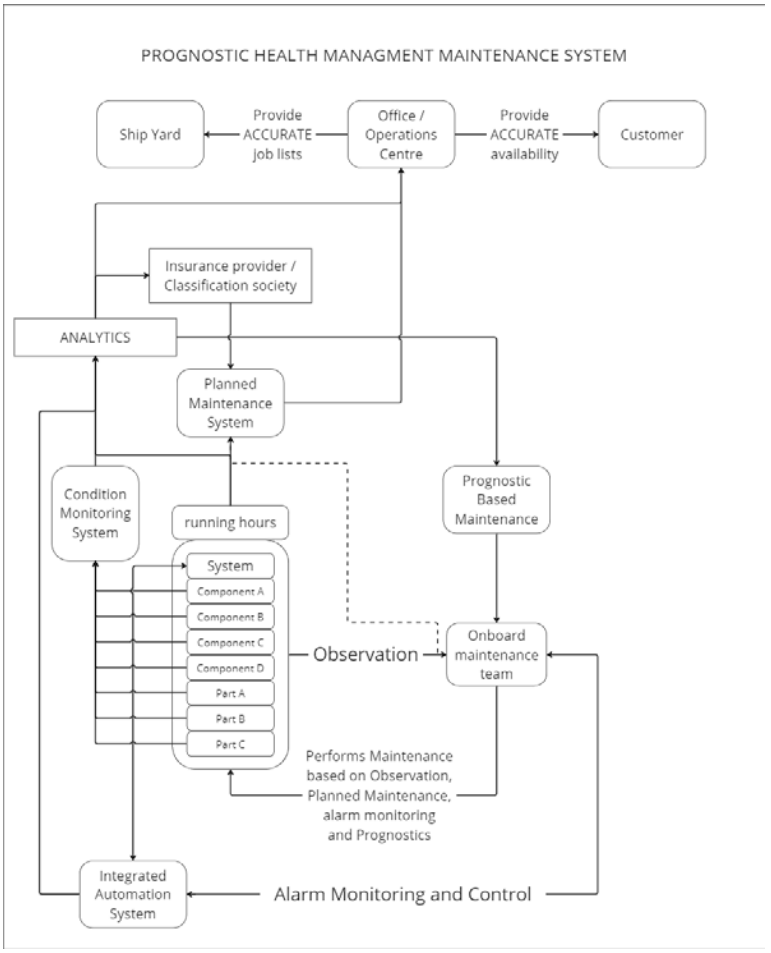


Figure 5 - Prognostic Health Management Maintenance System layout

9.1. Prognostic Health Management Maintenance System Description

We can see in figure 5 that the asset to be maintained is now broken into components and parts. Condition monitoring is applied to all components and parts. Covering the asset in monitoring sensors produces a lot of data. The monitoring system itself becomes complex and expensive. We can also see in figure 5 that the onboard maintenance team is relying on observations and prognostics only. Relying on total coverage for PHM eliminates the need for other maintenance strategies such as run to failure, but many failure modes, especially in the electronics domain, cannot be detected by readily available sensors.

9.2. Method of Performing Maintenance

As failure modes are detected action will be taken. In an ideal total PHM system all maintenance is done proactively and for data driven reasons

9.3. System Health Indicators

In this strategy the state of each failure mode for each system can be presented to the operator. For example, failure mode A may have not been detected, whereas failure mode B may be 10% of the way to critical failure of the system. As well as this the onboard team will use observations for system health indication.

Table 8: PHMMS metrics

KPI / SHI	Description
Availability	328 days per year potential
Human reliance / human error	5 humans obtaining information for maintenance
Unplanned maintenance tasks	2800 tasks
Planned maintenance tasks	2300 tasks
Set up cost	$1,000,000 – one time cost
Running costs	$680,000 – yearly running costs
Maintenance Costs	$500, 000 – yearly spares / consumables

Table 8 shows the average metrics for this maintenance system. For this particular strategy it is important to note the complexity of the monitoring system and the cost to set this up. At this level the monitoring system its self is likely to experience failures just due to the amount of sensors and the probability MTTF.

Due to complexity, cost, and the amount of time it would take to set this up.

10. RAM Enabled Predictive Maintenance System

We now look at the culmination or sweet spot system, a RAM enabled PHM system. The diagram below shows an example general layout of such a system.

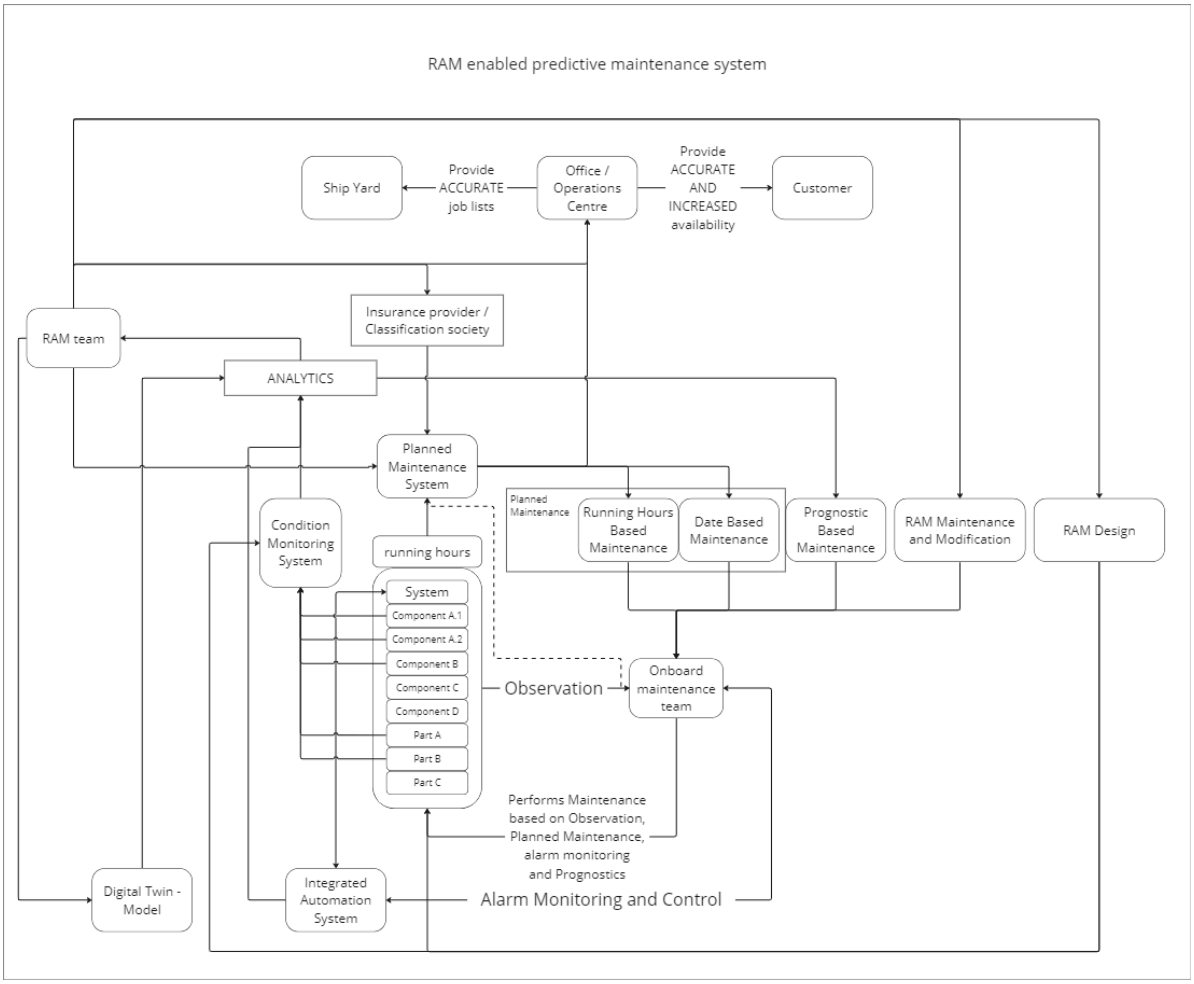


Figure 6. RAM enabled PHM Maintenance System layout

10.1. RAM Enabled Maintenance System Description

This system has many parts as can be seen in figure 6, but there is potential for harmony. The main difference between this and the previous system is in the design. Here targeting is done through RAM. No longer are we blanketing the asset in sensors, now we are using advanced design methods to create the minimum sensor set to cover the maximum failure modes. Digital risk and function twins are used to facilitate the sensor suit design and simulate effectiveness.

RAM shows us that certain components benefit from certain maintenance strategies depending on criticality, cost, redundancy and other factors.

10.1.1. Digital Twins

Following a concise RAM strategy generates details on the types of tools that can be used. Digital twins can play an important role in PHM design and assist in enabling effective PHM.

Digital twins are an aspect that is gaining momentum in the realms of PHM (Kammal Al-Kahwati, 2022) there are many advantages to using digital twins as part of machinery health management as Al-Kahwati explains There is an important point that is hinted at in Al-Kahwati's paper, that is availability. In order for the techniques in PHM to be given serious consideration by industry there must be a quantifiable gain. Availability of an asset is one such quantifiable metric, (the other two main areas being Reliability and Maintainability – RAM) Availability of a system is essential for a solid business case (Kammal Al-Kahwati, 2022). (Mulugeta Weldezgina Asres, 2022) AnoP is becoming increasingly linked to a concept known as Industry 4.0 (Mulugeta Weldezgina Asres, 2022) the ability to detect causal based anomalies of complex systems is critical to both the systems health and the quality of the system output. Using a multivariant causality-based anomaly prediction system as part of prognostic health management is about as advanced as system health prediction can get.

10.2. Method of Performing Maintenance

With this system the maintenance is still performed by the onboard maintenance team, however the maintenance is much more targeted. Spares holding can be reduced and potentially only ordered once degradation indicators are presented to the team. Systems and components deemed as non-critical and low cost are still replaced or repaired only when they fail, which in some cases is the most effective strategy. For example, light bulbs / tubes or LED's are run to failure items.

10.3. System Health Indicators

The health indicators are tuned per system. One system may only present human observable indicators, while another may present complete failure mode status through additional sensor sets. The indicators across the system of systems that is a vessel are optimised.

Table 9: RAM Enabled Predictive Maintenance System Metrics

KPI / SHI	Description
Availability	347 days per year potential
Human reliance / human error	<4 humans obtaining information for maintenance
Unplanned maintenance tasks	1000 tasks
Planned maintenance tasks	3000 tasks
Set up cost	$500,000 – one time cost
Running costs	$610,000 – yearly running costs
Maintenance Costs	$400, 000 – yearly spares / consumables

11. RAM / PHM Maintenance System For Lean / Uncrewed Vessels (MASS)

The last iteration of this maintenance system evolution is to tie the lean / uncrewed operational model to the RAM / PHM model.

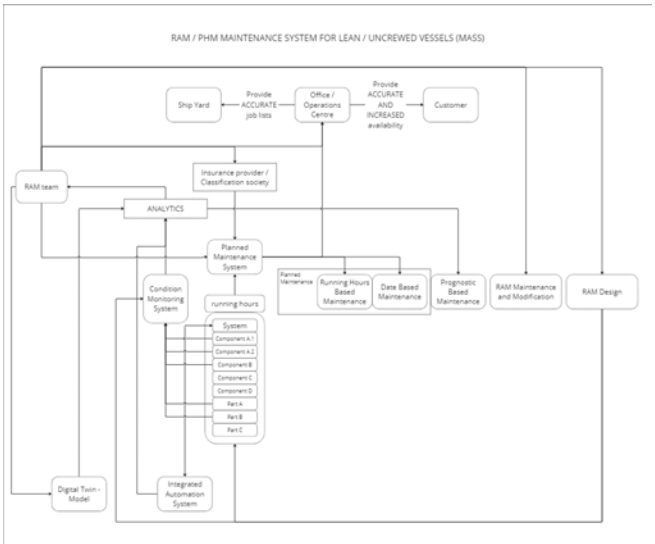


Figure 7. RAM / PHM system for MASS with onboard maintenance team removed

Figure 7 above shows the removal of the humans. Despite the sophisticated technology this removal leaves a gap between information and action. Figure 8 shows an example of how this gap may be filled by remote operations.

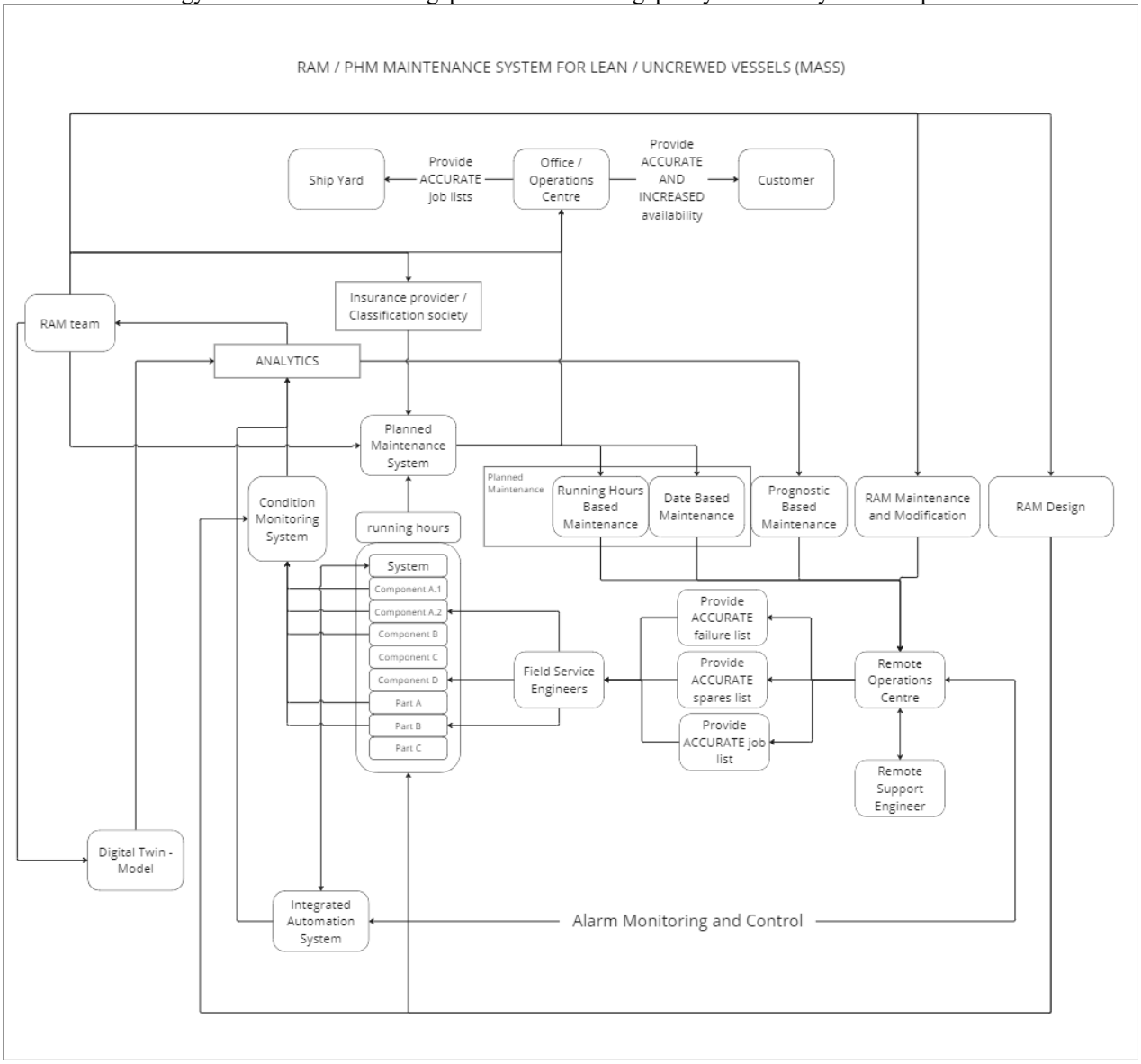


Figure 8. RAM / PHM system for MASS with onboard maintenance team replaced by remote operations

11.1. RAM / PHM Maintenance System For Lean / uncrewed Vessels (MASS) system description

The system its self is the same as the previous system, however the decision making has been moved to a shore based facility and the maintenance actions are now being performed by a specialist team that are only onboard when the vessel is in port. Figures 7 and 8 show that a RAM enabled PHM system facilitates lean / uncrewed operations while the need for lean / uncrewed operations validates the use of a RAM enabled PHM maintenance system.

11.2. Method of Performing Maintenance

With this system a shore side team are given instructions on what maintenance actions need to be done prior to joining the vessel. The list is completed and the team leave the vessel prior to her leaving port. These actions include those described with the previous system in section 10.

11.3. System Health Indicators

The same indicators are generated as in section 10 but without the human observations. Summary can be seen in table 10 below.

Table 10 : RAM / PHM Maintenance System for Lean / Uncrewed Vessels (MASS))

KPI / SHI	Description
Availability	356 days per year potential
Human reliance / human error	Potentially 0 humans obtaining information for maintenance
Unplanned maintenance tasks	Potentially 0 tasks
Planned maintenance tasks	3000 tasks
Set up cost	$600,000 – one time cost
Running costs	$360,000 – yearly running costs
Maintenance Costs	$300, 000 – yearly spares / consumables

12. DISCUSSIONS

A conventional Maintenance system is reliant on the maintenance crew onboard the vessel. The cost of the crew is high, and including the systems to support the crew onboard adds size and cost to the vessel. Diagnostics can be time consuming and there is a high amount of reactive maintenance. Having such a highly qualified maintenance crew onboard the vessel means the maintenance is self-managed onboard.

The general coverage of CMS and acceptance of PHM is a subject that can be heavily discussed. The latest numbers on CMS coverage are 14 years old, publications such as DNV's titled Beyond Condition Monitoring in the Maritime Industry is a fantastic snap shot of the state of CMS coverage around the same time as the coverage survey was conducted by Lloyds (Knut Erik Knutsen, 2014) does this suggest a new survey is needed?

It is also worth discussing the practical and theoretical implications to an asset and to an organisation if a RAM enabled PHM driven Maintenance system is employed. Practically for the asset there will be component changes to conform to the RAM strategy, spares holdings will change and additional sensor sets will be added. Systems that are not normally integrated may need to be integrated under a RAM / PHM maintenance system.

For the organisation there will need to be adjustments, both in the personnel skills and in the connections between departments. A new way of handling services will need to be developed including department and logistics handling to create the enhanced service team required to service an uncrewed vessel.

In theory, the gaps that an organisation faces and the gaps that the asset faces can be realised by proper assessment that takes into account RAM. This assessment must be carried out with appropriate subject matter experts in order to ensure the asset and the organisation are ready for the adoption of this new maintenance / operations aware design approach.

Table 11 below aims to summarize the KPI's and impact on them by the difference maintenance strategies. Taking into account the details from the processes presented in the previous sections. The cost figures associated with these KPI's were averaged from typical industry quotes related to maintenance upgrades and new build design. The costs will vary dependent on the vessel class, its anticipated modes of operation, and the age of the vessel if considering retro fit.

Table 11 : KPI / SHI summary

KPI	Metric / Unit	Conventional Maintenance System	Conventional Condition Based Maintenance System	Prognostic Health Management Maintenance System	RAM Enabled Predictive Maintenance System	RAM / PHM Maintenance System for Lean / Uncrewed Vessels (MASS)	Trend
Availability	days availability per year	325	325	350	350	358	
Human reliance / human error	maintenance staff needed onboard	5	5	5	4	0	
Unplanned maintenance tasks	average number of tasks per year	3200	3000	2800	1000	0	
Planned maintenance tasks	average number of tasks per year	2500	2800	2300	3000	3000	
Set up cost	average dollars one time cost	250000	300000	1000000	500000	600000	
Running costs	average dollars per year	680000	680000	680000	610000	360000	
Maintenance Costs	average dollars per year	500000	500000	500000	400000	300000	

Below are 3 consolidated graphs showing the general trend across the

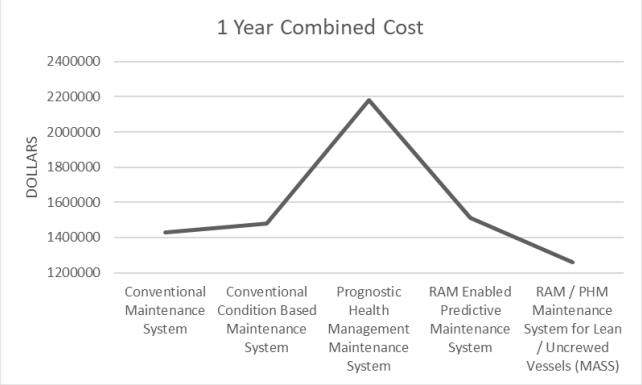


Figure 9. combined cost of each evolution over 1 year

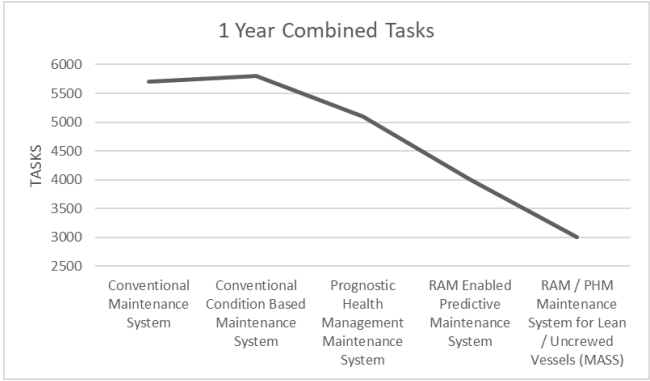


Figure 10. Combined tasks of each evolution over 1 year

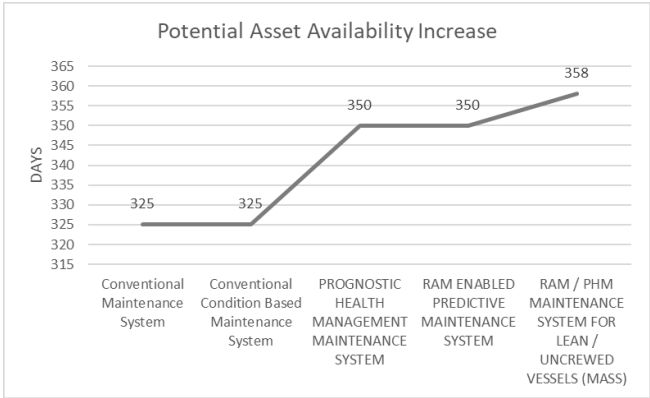


Figure 11. Potential asset availability using each evolution over 1 year

Figure 9 assists in demonstrating that a ROI is only feasible if RAM is included in PHM design.

Not using RAM could be a contributor to advanced monitoring system utilization being abandoned in many cases in the maritime industry due to a negative ROI

Figure 11 lets us consider the worth of an average day rate for the vessel. For example, take a day rate of $20,000, this gives a maximum of $7.3M generation. The income potential difference between a conventional maintenance system and a RAM enabled PHM maintenance system can be 10% which in this example equates to $660,000 per year additional potential income. If we take a conservative look at this and realise only 5% additional availability, the additional income will offset the cost of setting up a RAM enabled PHM system in 2 years. In addition to increased income potential there are decreases in costs that can be maximised by designing the vessel to be lean / uncrewed. There is also the increased reliability of the asset and increased availability prediction accuracy that can have major impacts on reputation.

While a RAM enabled PHM system is most effective when combined with the lean / uncrewed option it also enables the lean / uncrewed option and so these two technologies combined with their associated philosophies are bound together and are mutually beneficial.

When looking at the setup costs for PHM the following must be considered.

The PHM approach without using RAM is fraught with danger leading to high costs, high amounts of data. The extreme nature of this strategy could mean that components that were traditionally low cost run to failure items are now being monitored by an expensive and complex system. In this case maintenance / replacement costs will increase.

The design approach is another point of discussion, using digital twins and how to combine the data driven approach with subject matter experts to perfect the design. Extensive “tuning” must be done to ensure the best fit for each system onboard the vessel, this can be done much faster using digital twin technology.

Below is a generalised list of considerations for implementing a PHM system on any system onboard a vessel.

Table 3: PHM system components

SYSTEM	COMPONANT
Analysis	Software
Analysis	Engineer
Analysis	Training
Analysis	Computer
Analysis	Modelling
Data Acquisition	Sensors
Data Acquisition	Cabling
Data Acquisition	Cabinets
Data Acquisition	DAQ's
Data Acquisition	Servers
Data Acquisition	UPS
Data Acquisition	Transmission
Data Analytics	Purchasing Service
Data Analytics	Developing Service
Data Analytics	Training
Data Analytics	Engineer
Data Analytics	Storage

Table 11 shows an example of the components required to set up the proposed system, in addition the company itself must be setup to handle prognostics.

12.1. Implementation Examples

The proposed move to a RAM enabled PHM maintenance system should be employed if shipping owners have a need to increase asset availability. An implementation example would be a shipping company that is renewing / replacing vessels and wants the new vessels to have higher availability, reduced running costs, and remote capability, either lean or uncrewed.

Another example would be a vessel owner seeking to build more advanced uncrewed vessels and requiring a maintenance system that can facilitate the nature of uncrewed vessels.

12.2. Next Steps

The proposed move to RAM enabled maintenance systems can have wide reaching implications for the maritime industry, from ship builders and operators, to crew and service technicians, and then to assurance and insurance providers.

Ship builders can benefit from RAM enabled designs by offering increased reliability, operators share the same benefit. The crews serving onboard will be conducting maintenance in different ways. Assurance and insurance providers can benefit from the machinery health on demand that is achievable with the data produced by a RAM enabled PHM maintenance System. The future implications of the proposed system warrant extensive discussion in order to maximise the benefit to all stakeholders.

13. CONCLUSIONS

This paper describes a possible model for an advanced maintenance system that enables lean / uncrewed vessel operations. It also describes the evolutionary steps that have occurred to reach the proposed system.

Contributions to this paper are:

1. Articulation of maintenance strategies typically found in the maritime industry sector (including mapping of typical actives by stakeholders).
2. A snapshot of a generalised PHM value model targeted at the maritime industry sector.

One conclusion that can be seen is that a pure PHM approach is not effective and should be avoided. We can also conclude that the maritime industry is due for CMS coverage / utilisation / acceptance surveys, including acceptance of PHM from both engineering and cultural perspectives. The third conclusion from this paper is that there is a sweet spot for maintenance that can only be achieved by design, and that a concise RAM philosophy is an appropriate tool for assisting in the design and enabling a PHM system for vessels.

We can also conclude that there are existing elements onboard the vessel to build upon to facilitate lean / uncrewed operations, such as E0 and UMS notations.

The last conclusion is that a RAM enabled PHM maintenance system both supports and is validated by lean / uncrewed vessel operations and is a major contributor to asset availability increase.

14. ACKNOWLEDGEMENT

.I would like to acknowledge Simon Crompton, head of Quality Engineering at Ocean Infinity for supporting the writing of this paper.

15. NOMENCLATURE

AnoP - Anomaly Prediction
CBM - Condition Based Maintenance / (Condition Based Monitoring)
CM - Corrective Maintenance
CMS - Condition Monitoring System
DNV - Det Norske Veritas – (Risk Management & Quality Assurance)
ETO - Electro Technical Officer
IAS - Integrated Automation System
ISM - International Safety Management Code
IVHM - Integrated Vehicle Health Management
KPI - Key Performance Indicator
MAD - Maintenance Aware Design
MTTR - Mean Time To Repair
PHM - Prognostic Health Management
PM - Predictive Maintenance
PMS - Planned Maintenance System
PSA - Petroleum Safety Association
RAM - Reliability, Availability, Maintainability
RCM - Reliability Centred Maintenance
ROI - Return On Investment
SFI - The SFI Code is an international classification standard used in shipping.
SHI – System Health Indicator

16. REFERENCES

Alexander Athanasios Kamtsiuris, ,. F. (2022). A Health Index Framework for Condition Monitoring and Health Predictions. Turin, Italy.: Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Alexandre Trilla1, 3. N.-C. (2022). Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting. *International Journal of Prognostics and Health Management, ISSN2153-2648*.

Cornelius Scheffer, P. G. (2004). *Machinery Vibration Alanysis & Predictive Maintenance.*

Defeo, J. A. (2010). *Juran's Quality Handbook* (6th ed.). McGraw-Hill Professional.

H.J Hwang, J. L. (2018). A study of the development of a condition-based maintenance system for an LNG FPSO. *Ocean Engineering*.

Hyung Jun Park, ,. N.-H. (2022). A Comparative Study of Health Monitoring Sensors based on Prognostic Performance. Turin, Italy.: Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Kammal Al-Kahwati, W. B. (2022). Experiences of a Digital Twin Based Predictive Maintenance Solution for Belt Conveyor Systems. Turin, Italy.: Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Knut Erik Knutsen, G. M. (2014). *Beyond Condition Monitoring in the Maritime Industry.* DNV.

Maximilian-Peter Radtke1, J. B. (2022). Combining Knowledge and Deep Learning for Prognostics and Health Management. Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Miguel Simão1, R. P. (2022). Long-term Evaluation of the State-of-Health of Traction Lithium-ion Batteries in Operational Buses. *International Journal of Prognostics and Health Management, ISSN2153-2648*.

Mulugeta Weldezgina Asres, G. C. (2022). Long Horizon Anomaly Prediction in Multivariate Time Series with Causal Autoencoders. Turin, Italy.: Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Norway, S. (2010). Risk based maintenance & consequenceclassification. Lysaker, NORWAY: Standards Norway.

Qin Lang, K. E. (2024). A review of maritime equipment prognostics health management from a classification society perspective. *Ocean Engineering*.

Shipping, A. B. (2018). *Guidance Notes on SMART Function Implementation.* ABS.

Shorten, D. C. (2012). *Marine Machinery Condition Monitoring: Why has the shipping industry been slow to adopt?* Lloyd's Register London.

Sylvain Poupry, C. B. (2022). Towards data reliability based on triple redundancy and online outlier detection. Turin, Italy.: Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022 - ISBN – 978-1-936263-36-3.

Tilman Krokotsch1, M. K. (2022). Improving Semi-Supervised Learning for Remaining Useful Lifetime Estimation Through Self-Supervision. *International Journal of Prognostics and Health Management, ISSN2153-2648*.

17. BIOGRAPHIES

Toby Russell has 16 years at sea with the majority as chief ETO / Chief Electrical Engineer but also serving as 2nd engineer. Toby holds a MSc in applied Instrumentation and Control and he has a keen interest and passion for marine engineering that has progressed into the fields of RAM engineering to support the development of novel technologies to enable remote vessel operation on a commercial level. Toby has worked for Ocean Infinity for 2 years leading novel system development and integration projects as well as leading on development of RAM philosophies and strategies.

Octavian Niculita Octavian Niculita is a Senior Lecturer in Instrumentation with Glasgow Caledonian University. He has a PhD in Industrial Engineering from the Technical University of Iasi, Romania carried out under the EDSVS framework. His current research interests include industrial digitalisation, predictive maintenance, PHM system design, integration of PHM and asset design for aerospace, maritime, and oil & gas (surface and subsea) applications. Octav has over 15 years of experience in design and development of prognostics and health management applications, having worked on applied aerospace projects funded by The Boeing Company and BAE Systems as a Research Fellow and Technical Lead on his previous appointment with the IVHM Centre at Cranfield University, UK. He is a member of the Prognostics and Health Management Society, InstMC and the IET.

A Physics-Inspired and Data-Driven Approach for Temperature-Based Condition Monitoring

Giacomo Garegnani[1], Kai Hencken[1], and Frank Kassubek[1]

[1] *ABB Switzerland Ltd., Corporate Research*
giacomo.garegnani@ch.abb.com
kai.hencken@ch.abb.com
frank.kassubek@ch.abb.com

ABSTRACT

System overheating is a common problem in electric equipment, as degradation of contacts lead to an increase in Ohmic resistance and increased thermal losses. Temperature measurements are widely employed to monitor a device's health status, to estimate its remaining useful life, and to inform maintenance strategies. An issue that is special to electrical distribution networks is the varying heating power, which is in turn due to changes in the current. This results in varying temperatures, which in addition can often be delayed compared to the currents. Simple threshold-based diagnostics approaches are therefore not reliable for detecting faults in contacts. It is common to analyze the thermal behavior of electric devices using thermal networks, for both design and diagnostic purposes. Unfortunately, identifying the parameters of thermal networks from measured temperature data is a challenging problem, mainly due to identifiability issues and to numerical instabilities in parameter estimation. We propose an alternative data-driven strategy to compute the state-of-health of electrical devices which does not resort to thermal networks. Our approach consists in informing physics-based reduced models of the thermal response with sensor data. We show that our method is linked to the thermal network approach but avoids the full identification of the system, leading to better stability as well as less computational effort in the determination of its parameters. Rigorous testing with synthetic and experimental data confirms the efficacy of our methodology.

1. INTRODUCTION

The effective monitoring of the operational health of electric devices is of utmost importance to guarantee the secure and steady functioning of industrial facilities (Hoffmann et al., 2020). Among the vulnerabilities encountered by these devices, the issue of overheating due to Joule's effect stands out prominently. A significant portion of the heat generated within these devices comes from electric contacts. The deterioration of contacts results in an increase in their electrical contact resistance (ECR), which, in turn, triggers pronounced overheating. Such overheating not only disrupts operational stability but also exposes the devices to the imminent risk of irreparable harm.

With the rise in connectivity of industrial devices – the industrial internet of things – the potential of monitoring algorithms for predictive maintenance has grown considerably. In the present context, temperature data can be leveraged to prevent excessive overheating and monitor the health state of devices. While simple algorithms monitor the temperature and raise alerts based on critical levels, the dynamic nature of the thermal response to time-dependent current loads yields more insight into the root cause of the problem.

The method we propose in this report is computationally light and memory-efficient (in contrast to numerical solvers of partial differential equations), and is robust when confronted with real data (unlike thermal networks). Despite its simplicity, we believe this method can be effectively used for making thermal predictions and detect anomalous behavior for a wide range of electric devices.

Thermal networks can be cumbersome to set up and train (O. M. Brastein et al., 2019; O. Brastein et al., 2020; Boodi et al., 2022), but they are nevertheless a flexible tool to model the temperature response of the device. Conversely, the method which we propose here does not allow to predict temperatures away from the sensor and requires some dedicated training of the response of each device using a specific current profile. Memory-wise, the method also requires some limited storage of past current values. Finding a mathematical equivalence of thermal networks and the proposed method may therefore lead to the development of a method that combines the two approaches and retains the advantages of both.

The remainder of this article is as follows. In Section 2 we introduce the thermal kernel method, focusing on how to cope with noise in the data, and on how to infer variations in the ECR values. The equivalence of the kernel method and the thermal network approach is then discussed in Section 3. In Section 4 we test the method against synthetic data, checking the accuracy of the method as well as its robustness against model misspecification and noise in temperature data. Finally, in Section 5 we draw our conclusions and propose an outlook for future work.

2. THE THERMAL KERNEL METHOD

We consider an electric device which is equipped for simplicity with a single temperature sensor (see Section 2.4 below for the case of multiple sensors). We model the over-temperature T at the sensor location for a new and healthy device as

$$T(t) = \int_0^t k_0(t-s)\, I^2(s)\, \mathrm{d}s, \tag{1}$$

where k_0 is an unknown kernel function, and where I denotes the electric current, which we assume to satisfy $I(t) = 0$ for all $t \leq 0$. Note that k_0 can be seen as a Green function of the thermal problem, and I^2 is proportional to the thermal input due to Joule's law of heating. The kernel k_0 captures all linear thermal influences on the temperature measured by the sensor, and in particular the heat generated both by bulk conductors (e.g., busbars in electric devices) and the heat generated at imperfect contacts, which are both proportional to I^2. The existence of such a kernel function is guaranteed if we assume that the system is linear both with respect to the heat flow and the dependency on I^2.

Let N be the number of electrical contacts in the device, and denote by ΔR_i the variation (typically an increase) of the ECR in the i-th contact for $i = 1, \ldots, N$. Then, we assume that there exist kernels k_i for all $i = 0, 1, \ldots, N$ such that the temperature at the sensor location reads

$$\begin{aligned} T(t) = &\int_0^t k_0(t-s) I^2(s)\, \mathrm{d}s \\ &+ \sum_{i=1}^{N} \Delta R_i \int_0^t k_i(t-s) I^2(s)\, \mathrm{d}s. \end{aligned} \tag{2}$$

With the formula in Eq. (2), we assume that the thermal response at the sensor location of the device after contact degradation is encoded by the kernel

$$k_\Delta(t) = k_0(t) + \sum_{i=1}^{N} \Delta R_i k_i(t), \tag{3}$$

where the kernels $k_i(t)$ model the thermal response due to a change of resistance at contact i. The kernel functions k_i for $i = 0, 1, \ldots, N$ are unknown and depend on the device geometry, on how heat is exchanged with the surrounding environment, and on the thermal interconnections of the device components.

2.1. Determination of the kernel functions

In order to determine the kernel functions we use the response of the system to a step excitation, i.e., by imposing a constant current $I(t) = I_0$ for all $t \geq 0$. First, we fix $\Delta R_i = 0$ for all i and derive both sides of Eq. (1) with respect to t to obtain

$$k_0(t) = \frac{\dot{T}(t)}{I_0^2}. \tag{4}$$

Note that the temperature derivative $\dot{T}$ may not be available from sensor data, but can easily be reconstructed in practice by means of a finite difference formula from the measured temperature T. Note that numerical differentiation may amplify noise on the signal. We tackle this issue in Section 2.2 below. Given k_0 we can then determine the remaining N kernels by increasing the ECRs by a known quantity one by one. Indeed, if it holds $I(t) = I_0$ and $\Delta R_j = 0$ for all $j \neq i$ for a fixed index i, we have from Eq. (2)

$$\dot{T}(t) = \left(k_0(t) + \Delta R_i k_i(t)\right) I_0^2.$$

If the ECR increase ΔR_i is known and we measure the corresponding temperature T, then the kernel k_i is given by

$$k_i(t) = \frac{1}{I_0^2 \Delta R_i} \left(\dot{T}(t) - I_0^2 k_0(t)\right). \tag{5}$$

It might be unpractical or impossible in some scenarios to increase the ECR by a known quantity. Determining kernel functions may then involve data generation through a high fidelity simulation.

2.2. Noisy or short temperature data: Exponential fit

Let us assume that the temperature T is observed for a finite time interval $0 \leq t \leq t_{\mathrm{end}}$ and that observations are subject to measurement noise. In this case, the kernel k_0 given by Eq. (4) (and similarly the kernels k_i, $i = 1, \ldots, N$) should be post-processed to obtain a smooth kernel that can also be evaluated for times $t > t_{\mathrm{end}}$. For this purpose, we can introduce the natural assumption that k_0 is given by an infinite sum of negative exponential functions, as in

$$k_0(t) = \sum_{j=1}^{\infty} a_j \exp(-\lambda_j t),$$

where $a_j \in \mathbb{R}$, $\lambda_j \in \mathbb{R}^+$ for all $j = 1, 2, \ldots$. We then truncate the sum to an integer number N_{exp} of exponential functions and write

$$\widetilde{k}_0(t) = \sum_{j=1}^{N_{\mathrm{exp}}} a_j \exp(-\lambda_j t).$$

A suitable value $N_{\exp}$ can be chosen with a model selection algorithm. Finally, we determine a_j and λ_j by maximizing the likelihood of the noisy kernel k_0 given by Eq. (4) applied with the data sequence T. This ansatz can be motivated by the equivalence between thermal networks and kernels (or more in general by any finite-dimensional approximation of the full heat problem). More details can be found in Section 3.

The approach of fitting exponential functions to k_0 could be problematic in case temperature data are corrupted by noise. Indeed, noise is amplified when computing the time derivative $\dot{T}$ of the temperature. In this case, it is more robust to fit directly the temperature data T, which under the assumption above is approximated by

$$\widetilde{T}(t) = I_0^2 \sum_{j=1}^{N_{\exp}} \frac{a_j}{\lambda_j} \left(1 - \exp(-\lambda_j t)\right).$$

We can therefore fit the curve above directly to the temperature data and determine the values of a_j and λ_j which fully define the kernel function k_0.

We can repeat the same reasoning for the kernels $\{k_i\}_{i=1}^N$ modeling the thermal response at the sensor location due to (additional) heat generated at the contacts. We make the guess that for all $i = 1, \dots, N$ it holds

$$k_i(t) = \sum_{j=1}^{N_{\exp}} a_{ij} \exp(-\lambda_{ij} t).$$

Manipulating Eq. (5) with similar calculations as above we obtain

$$T(t) - I_0^2 \int_0^t k_0(t-s)\,\mathrm{d}s = I_0^2 \Delta R_i \sum_{j=1}^{N_{\exp}} \frac{a_{ij}}{\lambda_{ij}} \left(1 - \exp(-\lambda_{ij} t)\right).$$

The left-hand side of this equation is known. Fitting the coefficients a_{ij} and λ_{ij} to data then defines the kernel k_i. Note that this approach assumes that the coefficients λ are independent of each other for k_0 and each k_i. Since the thermal time scales should be the same for the nominal value of the resistance and increased resistances by linearity, the values λ_{ij} should be shared by the fit to k_0. A more robust approach, which we do not investigate here, would therefore consist in fitting the kernel functions simultaneously.

2.3. Inference of the resistance variations

In this section, we describe how knowledge of the kernel functions can be combined with temperature data to infer online a variation of the ECR of the N contacts, and consequently deduce their health status. Assume that all the kernels k_i have been determined and denote for $i = 0, \dots, N$ by K_i the integrated quantity

$$K_i(t) = \int_0^t k_i(t-s) I^2(s)\,\mathrm{d}s,$$

where I is the measured current. Then, we can rewrite Eq. (2) as

$$T(t) = K_0(t) + \sum_{i=1}^{N} \Delta R_i K_i(t). \tag{6}$$

Assume that the current and the temperature at the sensor have been measured on a set of times $\mathbf{t} = (t_0, t_1, \dots, t_M)$, where $t_j = t_s \cdot j$ and t_s is the sampling time. We can then assemble M-dimensional vectors

$$\mathbf{T} = T(\mathbf{t}), \qquad \mathbf{K}_i = K_i(\mathbf{t}),$$

where $T(\mathbf{t}) = (T(t_0), T(t_1), \dots, T(t_M))^\top$. Using the vectorial notation, the discrete version of Eq. (6) is

$$\mathbf{T} = \mathbf{K}_0 + \sum_{i=1}^{N} \Delta R_i \mathbf{K}_i.$$

An estimator $\widehat{\Delta R} \in \mathbb{R}^N$ of the vector of ECRs can be defined as the least square estimator

$$\widehat{\Delta R} = \arg\min_{\Delta R} \left\| \mathbf{K}\Delta R - (\mathbf{T} - \mathbf{K}_0) \right\|, \tag{7}$$

where $\mathbf{K}$ is the $M \times N$ matrix with columns $\mathbf{K}_i$ for $i = 1, \dots, N$. The minimization problem is overdetermined whenever $M \geq N$, i.e., the number of time instants for the measurements exceeds the number of contacts in the system, which is most likely verified. Hence, the estimator in Eq. (7) should be determined as the solution of the $N \times N$ linear system

$$\mathbf{K}^\top \mathbf{K} \widehat{\Delta R} = \mathbf{K}^\top (\mathbf{T} - \mathbf{K}_0).$$

Note that in real applications we expect the values of ΔR_i to increase rather than decrease due to contact degradation. A physically meaningful solution could therefore be enforced by using the constrained minimizer

$$\widehat{\Delta R} = \arg\min_{\Delta R \geq 0} \left\| \mathbf{K}\Delta R - (\mathbf{T} - \mathbf{K}_0) \right\|,$$

where the symbol $\geq$ is meant component-wise.

2.4. Multiple temperature sensors

We now consider a device which is equipped with multiple temperature sensors, and explain how more information can be leveraged to obtain a possibly more precise estimation of variations in the ECRs.

Assume that we have J temperature sensors. The temperature

of each sensor $j = 1, \dots, J$ can be expressed as

$$T^j(t) = \int_0^t k_0^j(t-s)I^2(s)\,\mathrm{d}s + \sum_{i=1}^{N} \Delta R_i \int_0^t k_i^j(t-s)I^2(s)\,\mathrm{d}s.$$

Note that the resistance increase ΔR_i is common for all sensors, as contacts are the same. Conversely, the temperature response is different across sensors, hence typically $k_i^{j_1} \neq k_i^{j_2}$ for $j_1 \neq j_2$. Kernels k_i^j can be determined as outlined in Section 2.1 for each $i = 0, \dots, N$ and $j = 1, \dots, J$. Similarly to Section 2.3, we then write $\mathbf{T}^j = T^j(\mathbf{t})$ and $\mathbf{K}_i^j = K_i^j(\mathbf{t})$ where

$$K_i^j(t) = \int_0^t k_i^j(t-s)I^2(s)\,\mathrm{d}s.$$

Calling $\mathbf{K}^j$ the $N \times M$ matrix whose columns are the vectors $\mathbf{K}_i^j$ for $i = 1, \dots, N$, we have J linear equations for ΔR

$$\mathbf{K}^j \Delta R = \mathbf{T}^j - \mathbf{K}_0^j, \qquad j = 1, \dots, J.$$

In order to compute the least square solution $\widehat{\Delta R}$ we assemble a $NJ \times M$ matrix $\mathbf{K}$ and NJ vectors $\mathbf{T}$ and $\mathbf{K}_0$ by stacking vertically the J equations as

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}^1 \\ \mathbf{K}^2 \\ \vdots \\ \mathbf{K}^J \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} \mathbf{T}^1 \\ \mathbf{T}^2 \\ \vdots \\ \mathbf{T}^J \end{pmatrix}, \quad \mathbf{K}_0 = \begin{pmatrix} \mathbf{K}_0^1 \\ \mathbf{K}_0^2 \\ \vdots \\ \mathbf{K}_0^J \end{pmatrix}.$$

The least square estimate $\widehat{\Delta R}$ is then the solution of the $N \times N$ linear system

$$\mathbf{K}^\top \mathbf{K} \widehat{\Delta R} = \mathbf{K}^\top (\mathbf{T} - \mathbf{K}_0),$$

and similarly to the single-sensor case a non-negative constraint can be imposed to the least-square solution. We note that in this case we expect an improvement by enforcing the time scale parameters to be the same across resistances when performing an exponential fit as in Section 2.2.

3. Equivalence with thermal networks

Thermal networks have been used to model the temperature of electric devices, and to infer health status given temperature measurements (Stosur et al., 2016). In this section, we describe how our approach simplifies thermal networks, whose parameters are notoriously difficult to estimate from data (O. M. Brastein et al., 2019; O. Brastein et al., 2020; Boodi et al., 2022). For a general discussion on identifiability of linear models, we refer the reader to (Raue et al., 2014).

We call thermal network a model which splits the device into an integer number N_c of compartments, whose temperature is assumed to be sufficiently homogeneous to be described by a single over-temperature T_i, for $i = 1, \dots, N_c$. We assume that the i-th compartment has a heat capacity C_i for $i = 1, \dots, N_c$. The compartments are thermally interconnected so that the heat flowing between the compartments indexed by i and j is proportional to their temperature difference with a constant h_{ij}. If two compartments are not directly connected thermally, we trivially set $h_{ij} = 0$. Moreover, we assume that the heat flowing towards the environment is proportional to the over-temperature T_i with a constant α_i. Finally, we assume that all elements in the network represent parts of the device which are subject to an electrical current $I = I(t)$, so that the thermal input to the i-th element is given by $u_i(t) = R_i I^2(t)$ by Ohmic heating. Under these assumptions, the over-temperature T_i of the i-th compartment of the network, for $i = 1, \dots, N_c$, satisfies the ordinary differential equation (ODE)

$$C_i \dot{T}_i(t) = \sum_{j=1, j\neq i}^{N_c} h_{ij}(T_j - T_i) - \alpha_i T_i + u_i(t). \tag{8}$$

In this section, we show how the temperature evolution of each compartment in a thermal network satisfies Eq. (1), i.e., there exist kernels k_0^i such that

$$T_i(t) = \int_0^t k_0^i(t-s)I^2(s)\,\mathrm{d}s, \tag{9}$$

for each $i = 1, \dots, N_c$, and that the kernel function can be written as a sum of exponential functions as in Section 2.2. Hence, a system whose thermal response can be described accurately by a thermal network can also be described by thermal kernels, with the advantage that in the kernel approach less parameters need to be determined from temperature measurements.

To start the derivation, we notice that the ODE system Eq. (8) can be written in matrix form as

$$\mathbf{C}\dot{\mathbf{T}}(t) = \mathbf{H}\mathbf{T}(t) + \mathbf{R}I^2(t), \tag{10}$$

where $\mathbf{T}$ is a vector with the temperatures of all compartments, where $\mathbf{R}$ is a N_c-dimensional vector containing the values of the resistances, and where $\mathbf{C}$ and $\mathbf{H}$ are appropriate matrices containing the values of the coefficients h, α and C. Let us rewrite Eq. (10) as the generic linear system

$$\dot{\mathbf{T}}(t) = -\mathbf{A}\mathbf{T}(t) + \mathbf{F}(t), \tag{11}$$

where $\mathbf{A} = -\mathbf{C}^{-1}\mathbf{H}$ and $\mathbf{F}(t) = \mathbf{C}^{-1}\mathbf{R}I^2(t)$. Let $\mathbf{T}(0) = \mathbf{T}_0$ be a known initial condition. It is simple to verify by differentiation that the solution of Eq. (11) is given by

$$\mathbf{T}(t) = \mathbf{e}^{-\mathbf{A}t}\mathbf{T}_0 + \int_0^t \mathbf{e}^{-\mathbf{A}(t-s)}\mathbf{F}(s)\,\mathrm{d}s, \tag{12}$$

where we denote by $\mathbf{e}^{-\mathbf{A}t}$ the matrix exponential to distin-

guish it from the scalar exponential (e.g., e^t).

The matrix $\mathbf{A}$ is not symmetric but it is diagonalizable with real eigenpairs.[1] Recall that for any diagonalizable matrix $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{V}$ is the matrix with the eigenvectors $\{\mathbf{v}_j\}_{j=1}^{N_c}$ of $\mathbf{A}$ as columns, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_{N_c})$ is the matrix of the eigenvalues, it holds

$$\mathbf{e}^{-\mathbf{A}t} = \mathbf{V}\mathbf{e}^{-\mathbf{\Lambda}t}\mathbf{V}^{-1} = \mathbf{V}\mathrm{diag}\left(e^{-\lambda_1 t}, \dots, e^{-\lambda_{N_c} t}\right)\mathbf{V}^{-1}.$$

This implies that if (λ, v) is an eigenpair of $\mathbf{A}$, then $(e^{-\lambda t}, v)$ is an eigenpair of $\mathbf{e}^{-\mathbf{A}t}$. Let $\mathbf{w}$ be an arbitrary vector in $\mathbb{R}^{N_c}$ and let $\{c_j = (\mathbf{V}^{-1}\mathbf{w})_j\}_{j=1}^{N_c}$ be the components[2] of $\mathbf{w}$ in the basis formed by the eigenvectors of $\mathbf{A}$, i.e., the scalars such that

$$\mathbf{w} = \sum_{j=1}^{N_c} c_j \mathbf{v}_j.$$

Hence, applying $\mathbf{e}^{-\mathbf{A}t}$ to $\mathbf{w}$ gives

$$\mathbf{e}^{-\mathbf{A}t}\mathbf{w} = \sum_{j=1}^{N_c} c_j \mathbf{e}^{-\mathbf{A}t}\mathbf{v}_j = \sum_{j=1}^{N_c} c_j e^{-\lambda_j t}\mathbf{v}_j.$$

Assume for simplicity and without loss of generality that $\mathbf{T}_0 = 0$. Replacing the decomposition above into Eq. (12) with $\mathbf{w} = \mathbf{C}^{-1}\mathbf{R}I^2(s)$ shows that

$$\mathbf{T}(t) = \int_0^t \sum_{j=1}^{N_c} c_j \mathbf{v}_j e^{-\lambda_j (t-s)} I^2(s)\,\mathrm{d}s,$$

where $c_j = (\mathbf{V}^{-1}\mathbf{C}^{-1}\mathbf{R})_j$. Hence, the temperature of the i-th compartment satisfies

$$T_i(t) = \int_0^t \sum_{j=1}^{N_c} \alpha_{ij} e^{-\lambda_j (t-s)} I^2(s)\,\mathrm{d}s,$$

where $\alpha_{ij} = \mathbf{V}_{ij} c_j$. This shows that the temperature of the i-th compartment can be indeed written as in Eq. (9) for

$$k_0^i(t) = \sum_{j=1}^{N_c} \alpha_{ij} e^{-\lambda_j t},$$

which is a sum of exponential functions as the approximations we employ in Section 2.2. Consider now that for each $k = 1, \dots, N_c$ the resistance of the k-th compartment increases by a quantity ΔR_k. We can write the overall kernel defining the temperature of the i-th compartment as

$$k^i(t) = \sum_{j,k=1}^{N_c} \mathbf{V}_{ij}\left(\mathbf{V}^{-1}\mathbf{C}^{-1}\right)_{jk} R_k e^{-\lambda_j t} + \sum_{j,k=1}^{N_c} \mathbf{V}_{ij}\left(\mathbf{V}^{-1}\mathbf{C}^{-1}\right)_{jk} \Delta R_k e^{-\lambda_j t}.$$

We see that k^i has the form of the kernel of Eq. (3) with

$$k_k^i(t) \coloneqq \sum_{j=1}^{N_c} \mathbf{V}_{ij}\left(\mathbf{V}^{-1}\right)_{jk} e^{-\lambda_j t},$$

which is the kernel associated to an increase in the k-th resistance as seen by the i-th element of the thermal network. Note that since the resistances do not appear in the expression of the system matrix $\mathbf{A} = -\mathbf{C}^{-1}\mathbf{H}$, the time scales λ_j in the kernels k_k^i are the same as the ones of the original kernel.

3.1. Generalization: kernel structure of thermal problems

The considerations above for thermal networks and the kernel structure of their solution applies more widely. In a linear approximation, heat transfer can be described by

$$\mathbf{C}\dot{\mathbf{T}} = \mathbf{L}\mathbf{T} + u, \tag{13}$$

where $\mathbf{L}$ is an operator describing both heat conduction $\mathbf{H}$ and coupling to the ambient α, and u is the heat injected in the system. In the specific case of a thermal network, the temperatures are vectors and the operators (finite dimensional) matrices. However, this equation may also describe a temperature field with a partial differential operator describing heat conduction on a physical domain Ω. For $x \in \Omega$, the local operator $\mathbf{C} = \mathbf{C}(x)$ is the specific heat capacity and the differential operator $\mathbf{L}(x) = -\nabla k(x) \cdot \nabla - k(x)\Delta$ describes heat conduction with a space-dependent heat conductivity k defined on Ω.

Note that Eq. (13) is linear in temperature, the operator $\mathbf{L}$ is self adjoint due to the symmetric nature of heat diffusion, and the field $\mathbf{C}$ is a (local) positive scalar. Normalizing the temperature $\widetilde{\mathbf{T}} = \mathbf{C}^{1/2}\mathbf{T}$ and multiplying Eq. (13) by $\mathbf{C}^{-1/2}$, we see that the operator occurring on the right hand side of the equation for $\widetilde{\mathbf{T}}$ ($\mathbf{C}^{-1/2}\mathbf{L}\mathbf{C}^{-1/2}$) is also self-adjoint. The spectral theorem then guarantees that this operator has real eigenvalues and orthogonal eigenfunctions that span the full space. Formally, the solution can be expressed in terms of the exponential operator $\mathbf{e}^{\mathbf{L}t}$ as

$$\widetilde{\mathbf{T}} = \int_0^t \mathbf{e}^{\mathbf{L}(t-s)}\widetilde{u}(s)\,\mathrm{d}s, \tag{14}$$

[1] Since $\mathbf{A} = -\mathbf{C}^{-1}\mathbf{H}$, with $\mathbf{H}$ symmetric and $\mathbf{C}$ diagonal and positive definite, we can write

$$\mathbf{A} = \mathbf{C}^{-1/2}\widetilde{\mathbf{A}}\mathbf{C}^{1/2},$$

where $\widetilde{\mathbf{A}} = -\mathbf{C}^{-1/2}\mathbf{H}\mathbf{C}^{-1/2}$. The matrix $\widetilde{\mathbf{A}}$ is real and symmetric, and hence can be diagonalized with real eigenpairs, which in turn implies that $\mathbf{A}$ is diagonalizable with real eigenpairs.

[2] Since the matrix $\mathbf{A}$ is in general not symmetric, the vectors $\mathbf{V}$ do not form a orthonormal basis of $\mathbb{R}^{N_c}$. If $\mathbf{A}$ is symmetric, it holds $\mathbf{V}^{-1} = \mathbf{V}^\top$ and

$$c_j = (\mathbf{V}^{-1}\mathbf{w})_j = \sum_{i=1}^{N_c} (\mathbf{V}^\top)_{ji}\mathbf{w}_i = \sum_{i=1}^{N_c} \mathbf{V}_{ij}\mathbf{w}_i = \langle \mathbf{v}_j, \mathbf{w}\rangle,$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean scalar product, which gives the more recognizable decomposition on a basis of orthonormal eigenvectors.

where $\widetilde{u} = \mathbf{C}^{-1/2}u$. For practical calculations, one has to expand in the eigenvectors as shown in the explicit example above. The general solution (14) has the same structure as the thermal kernels (1), which is hence a generic form for this type of linear heat diffusion problems. Therefore the exponential form of the kernel function can be derived independent of the assumption of an underlying thermal network as an approximation taking the dominant eigenmodes of $\mathbf{L}$ into account.

4. Numerical Experiments

In this section, we present a series of numerical experiments demonstrating the usefulness, accuracy, and robustness of our approach.

4.1. Scenario 1: Simple network

The first test setup we employ in experiments is represented schematically in Fig. 1(a). We consider an electrical device, e.g., a power protection device such as a breaker or a switch, which protects an electrical installation. The device connects the installation to a power source (e.g., the grid) with two electric contacts between busbars, one per side of the device. We assume that the device is equipped with a temperature sensor. We consider the problem of monitoring the ECR of the two contacts using the temperature sensor of the device.

In order to simulate this scenario, we use a three-compartments thermal network as shown in Fig. 1(b). In the network, the center element represents the device, and the lateral elements the two contacts. We suppose that the three compartments are exposed to the same ambient temperature T_{amb}, which we assume without loss of generality to be equal to zero.

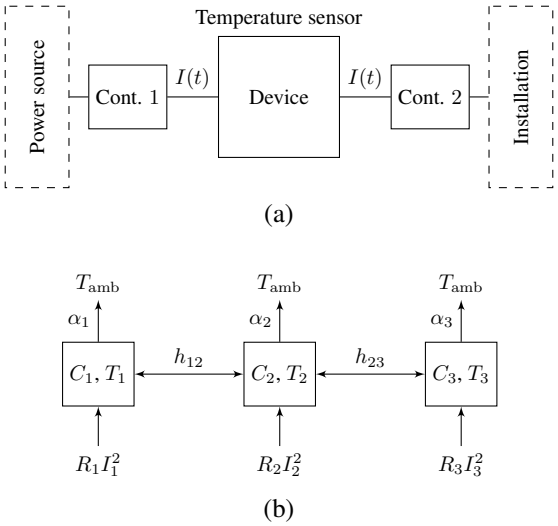


Figure 1. Setup for numerical experiments. (a) Schematic representation of a power protection device connecting an installation to a power source with two electric contacts. (b) Thermal network used to simulate the scenario.

The values of the coefficients C_i, α_i, R_i for $i = 1, \ldots, 3$, as well as of the h_{ij} for $(i, j) \in \{(1, 2), (2, 3)\}$, given in Table 1, are fixed to values which are realistic for a typical electric device. We determine the base kernel k_0 associated with the temperature sensor placed on the device fixing $I = 1\,\mathrm{kA}$ and simulating the network temperatures for $0 \leq t \leq 5\,\mathrm{h}$. Simulated data are obtained with an implicit numerical discretization of Eq. (10) on a time grid with time step equal to $1\,\mathrm{min}$. We then extract the device temperature T_2 and compute k_0 using Eq. (4), where $\dot{T}_2$ is computed by finite differences. We determine the kernels k_1 and k_3 associated to an increase of R_1 and R_3 following the procedure outlined in Section 2.1 with $\Delta R_i = R_i$, i.e., we double the ECR value to determine the kernel associated to a fault in the i-th contact.

Table 1. Coefficients of the thermal network in Fig. 1.

	$\alpha\,[\mathrm{W\,K^{-1}}]$	$R\,[\mu\Omega]$	$C\,[\mathrm{J\,K^{-1}}]$	$h\,[\mathrm{W\,K^{-1}}]$
1	1.0	100	3500	–
2	2.0	50	3500	–
3	3.0	100	3500	–
12	–	–	–	0.75
23	–	–	–	0.55

We measure the error on the i-th resistance as

$$\mathrm{err}_i = \frac{\left|\widehat{\Delta R}_i - \Delta R_i\right|}{R_i + \Delta R_i}, \tag{15}$$

where $\widehat{\Delta R}_i$ is the inferred increase in resistance and R_i is the nominal value of the i-th resistance (i.e., before increase). Note that the numerator in the right-hand side of Eq. (15) is equal to $|R_i + \widehat{\Delta R}_i - (R_i + \Delta R_i)|$, i.e., the absolute difference between the increased resistance and its inferred value. Hence, the error metric above is a relative error between the inferred and the true values of the increased *resistance*, rather than the resistance increase.

We generate 200 values of resistance increases $(\Delta R_1, \Delta R_3)$ randomly as $\Delta R_i \sim \mathcal{U}(0, R_i)$, independently for $i = 1, 3$. This means that the ECR degrades in all experiments, with values up to twice the original. For each pair of increases in the resistances, we generate 12 hours of temperature T_2 with sampling time $1\,\mathrm{min}$, always with the same current I defined by

$$I(t) = \begin{cases} 1\,\mathrm{kA}, & t \leq 1\,\mathrm{h}, \\ 0\,\mathrm{kA}, & 2\,\mathrm{h} < t \leq 5\,\mathrm{h}, \\ 0.7\,\mathrm{kA}, & 5\,\mathrm{h} < t \leq 9\,\mathrm{h}, \\ 0.3\,\mathrm{kA}, & t > 9\,\mathrm{h}. \end{cases} \tag{16}$$

The error in the estimation procedure, computed using Eq. (15), is summarized with boxplots in Fig. 2(a). We see that both resistances are estimated very accurately over the whole dataset of 200 experiments. Specifically, the error on R_1 never exceeds 0.1%, and the error on R_3 never exceeds 1%.

We repeat the same experiment but increase either R_1 or R_3 while keeping the other resistance to its nominal value. We repeat the inference 200 times per resistance with random increments as above. This experiment is relevant for applications, as the ECR of one contact only could undergo a rapid degradation, while all others could stay constant. Results, given in Figs. 2(b) and 2(c), demonstrates that also in this case the inference procedure is very accurate in determining the increased resistance values.

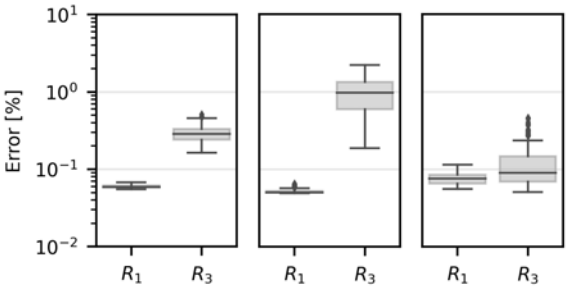


Figure 2. Percentage relative error in inference of two increased resistances given one temperature sensor. (a) Simultaneous increase of R_1 and R_3. (b) Only R_1 is increased. (c) Only R_3 is increased. Model configurations given in Fig. 1, test setup given in Section 4.1. Box-plot whiskers indicate 1.5 times the interquartile range, dots indicate outliers.

4.2. Scenario 2: The impact of adding a sensor

We consider a more complex configuration consisting of a thermal network with 6 compartments connected on a line, i.e., such that $h_{ij} = 0$ if $j \notin \{i-1, i+1\}$ for $i, j = 1, \dots, 6$. We assume that all parameters appearing in Eq. (8) are known, including nominal resistance values. Nominal parameter values are of the same magnitude as those of Table 1. Similarly to Section 4.1, we then increase randomly the resistances up to double their value and infer the increase increase with the procedure described in Section 2.3. The current used to excite the network with increased resistances is given in Eq. (16). We compare results obtained observing one temperature of the network only, T_2, and with two temperatures, T_2 and T_6. Note that when we observe one temperature we have one kernel k_0 and 6 additional kernels for the increase of R_i, $i = 1, \dots, 6$. When we observe two temperatures, we have one base kernel per sensor, and 6 additional kernels corresponding to an increase in resistance per sensor, for a total of 14 kernel functions. We recall that the method to infer the resistance increase with multiple sensors is described in Section 2.4.

Results, given in Fig. 3, demonstrate that errors can be as high as 60% on the fifth and sixth resistance (using the metric of Eq. (15)) when only the temperature R_2 is measured. This is because the thermal impact of the sixth compartment on the second is weak, and diluted by heat diffusion through the network. If we observe both T_2 and T_6, the error on all resistances is extremely low in most cases (below 0.01%), except of some outliers for which the error is above 50% error on R_6. This experiment nevertheless shows the benefit of equipping an electric device with an additional temperature sensor, especially if the device consists of many components that are thermally interconnected.

4.3. Scenario 3: Model misspecification

The method we present in this report to determine contact resistances relies on accurate determination of the kernel functions k_0 and k_i for $i = 1, \dots, N$. In a realistic setting, the kernel k_0 can be simply determined by applying a step current and measuring the temperature increase, or with any other system identification approach using data measured on the real device. For the kernels k_i, instead, we would need to increase artificially each resistance by a known quantity before applying a step current. It could be difficult, or unfeasible, to obtain such a controlled increase in practice, especially in a device-specific fashion. We could instead determine kernels that fit an entire fleet of devices, modulo the variability due to different installations. Specifically, we could use an experimental or simulated setup to determine universal resistance kernels $\widetilde{k}_i$ that are common to a whole fleet of devices, maintaining a base kernel k_0 that is specific to an individual installed device. The inferred resistances are then obtained as the solution to the linear system

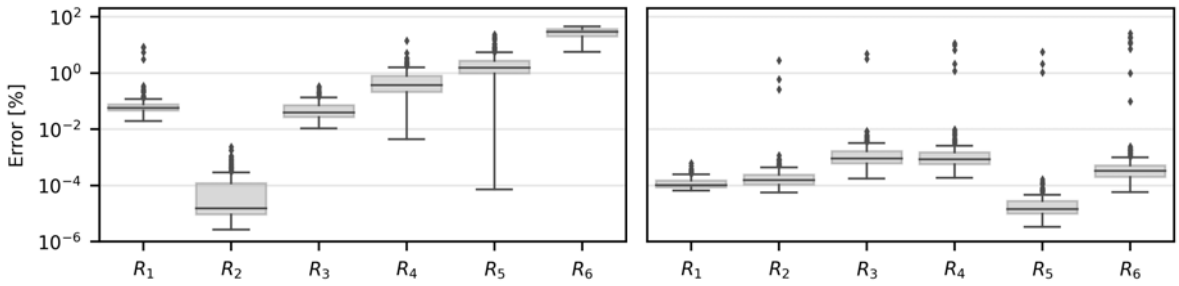
$$\widetilde{\mathbf{K}}^\top \widetilde{\mathbf{K}} \widehat{\Delta R} = \widetilde{\mathbf{K}}^\top (\mathbf{T} - \mathbf{K}_0), \tag{17}$$

where $\widetilde{\mathbf{K}}$ is built as in Section 2.3 using the fixed, universal kernels $\widetilde{k}_i$. The major concern with this approach is the misspecification between the real kernel function k_0 and the universal ones k_i, especially in terms of incompatible time scales.

Summarizing, the procedure that we propose to deal with installation specificity would consist of the following steps:

- Determine a universal base kernel $\widetilde{k}_0$ in an experimental or simulated setup;
- Use $\widetilde{k}_0$ to determine universal kernels $\widetilde{k}_i$ for each resistance that needs to be monitored;
- For each installation of the device, redetermine device-specific base kernel k_0 applying constant current load;
- When needed, infer an increase in resistances using Eq. (17).

We test the procedure above using the three-compartment network of Fig. 1, with coefficients given in Table 1. In order to simulate installation-specific conditions, we modify multiple times the value of the coefficients α_i as $\widetilde{\alpha}_i \sim \log \mathcal{N}(\alpha_i, \sigma)$, for $i = 1, \dots, 3$, where a large value of σ mimics devices that are very sensitive to different installations. We consider



(a) Compartment 2 sensed (b) Compartments 2 and 5 sensed

Figure 3. Inference of five resistances varied simultaneously in a six-compartment network. (a) One temperature sensor. (b) Two temperature sensors. Test setup given in Section 4.2. Box-plot whiskers indicate 1.5 times the interquartile range, and dots indicate outliers.

$\sigma = 0.4, 0.2, 0.1, 0.05$, and for each of these values we generate 200 values at random of the coefficients α to simulate 200 installations of the same device. Then, we infer the resistances R_1 and R_3 using Eq. (17). Note that we do not apply a resistance increase in this case, and just attempt to infer how impactful is a change of the nominal conditions onto the kernels.

Results, given in Fig. 4, demonstrate that re-calibrating only the base kernel k_0 for each installation is sufficient for keeping good accuracy in the inference of the resistances. Moreover, we see that a good inference result (below 1% except some outliers) can be achieved even for devices that are subject to high variability when installed (see the spread in temperature development in case $\sigma = 0.4$). We note that lower installation specificity results in smaller variability in the inferred resistances (see the width of the box-plots in case $\sigma = 0.05$).

4.4. Scenario 4: Noisy data

In all experiments above, we employed noiseless data for determining the kernel functions and for estimating the resistances in the model. In this section, we assess the impact of these two sources of noise on the estimation variability. We consider the simple three-element network of Fig. 1 with parameters as in the experiments above. We compute the base kernel k_0 and the kernels k_i associated to resistances $i = 1, 3$ by perturbing the temperature response to a step current with a Gaussian source of noise $\eta_k \sim \mathcal{N}(0, \sigma_k^2)$, where $\sigma_k > 0$. Then, we excite the system with the current profile of Eq. (16) and perturb the temperature response with a Gaussian source of noise $\eta_d \sim \mathcal{N}(0, \sigma_d^2)$, where $\sigma_d > 0$. We then infer the resistance increase without changing its value in the model, i.e., data is generated by imposing $\Delta R_i = 0$. We repeat the experiment for noise scales σ_d and σ_k ranging between 10^{-4} and 10^{-1}, and for each combination of σ_k and σ_d we repeat the experiment $M = 50$ times. At each j-th repetition, we record the estimated resistance variations $\widehat{\Delta R}_1^{(j)}$ and $\widehat{\Delta R}_3^{(j)}$. We measure variability in the estimation as the sum of the population standard deviations of the two estimated resistance increases, i.e.,

$$\text{variability} = \text{std}\left(\left\{\widehat{\Delta R}_1^{(j)}\right\}_{j=1}^{M}\right) + \text{std}\left(\left\{\widehat{\Delta R}_3^{(j)}\right\}_{j=1}^{M}\right),$$

where $\text{std}(\cdot)$ denotes population standard deviation. We repeat the estimation twice: once with raw kernel functions computed with Eqs. (4) and (5), and once by fitting exponential functions as explained in Section 4.4

Results, given in Fig. 5, demonstrate that the method we develop here is robust with respect to random sources of noise. As expected, the variability is a growing function of both σ_k and σ_d. We remark that fitting exponential functions to the kernels has a beneficial effect on the inference accuracy. Indeed, it can be noticed in Fig. 5 that the variability is slightly lower when thermal kernels are fitted with exponential functions.

5. Conclusion

We introduced a novel method based on thermal kernels to monitor the condition of an electric device given temperature measurements. This method allows the calculation of temperatures at specific locations for general linear heat diffusion problems including thermal networks, for which we demonstrated an equivalence analytically.

Thermal kernels are simple to fit to data due to their non-parametric nature, which prevents issues of poor identifiability. Indeed, the parameters of thermal networks as simple as the one of Fig. 1 can be cumbersome or even impossible to determine if only one of the compartments equips a temperature sensor, unless good priors on the parameters are available due to physical considerations or from the results of high-fidelity and high-cost simulations. This issue is completely circumvented by thermal networks, which absorb the effects of all the parameters of an equivalent network approach into

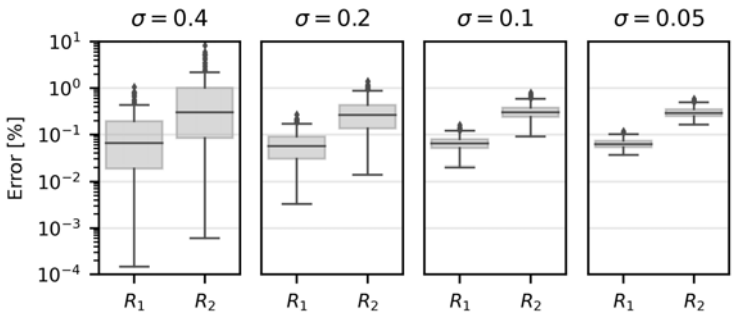
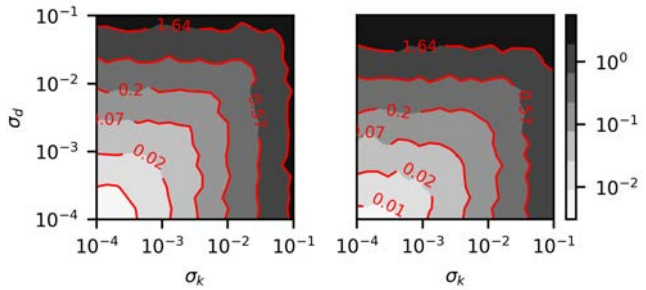


Figure 4. Results for four scales of model misspecification σ. Experiment setup in Section 4.3.



(a) Fitted kernels (b) Raw kernels

Figure 5. Estimation variability as a function of noise in the determination of the kernel functions (σ_k, horizontal axis), and in the data used for estimating the resistance values (σ_d, vertical axis). The contour values are in μΩ (a): Exponential fit for the kernel functions. (b): Raw kernel functions. Experiment setup in Section 4.4.

a simple data-driven linear transfer function.

We believe that thermal kernels should be preferred to thermal networks to monitor the linear heat sources (e.g., electrical resistances) of devices that are not equipped with a multitude of sensors, which would be required to fit the parameters of the network.

We suggest that future investigation may exploit the equivalence of thermal kernels and networks, e.g., to study whether knowing the former can be beneficial to improve the identifiability of the latter.

REFERENCES

Boodi, A., Beddiar, K., Amirat, Y., & Benbouzid, M. (2022). Building thermal-network models: a comparative analysis, recommendations, and perspectives. *Energies*, *15*(4), 1328.

Brastein, O., Ghaderi, A., Pfeiffer, C., & Skeie, N.-O. (2020). Analysing uncertainty in parameter estimation and prediction for grey-box building thermal behaviour models. *Energy and Buildings*, *224*, 110236.

Brastein, O. M., Lie, B., Sharma, R., & Skeie, N.-O. (2019). Parameter estimation for externally simulated thermal network models. *Energy and Buildings*, *191*, 200–210.

Hoffmann, M. W., Wildermuth, S., Gitzel, R., Boyaci, A., Gebhardt, J., Kaul, H., ... Tornede, T. (2020). Integration of novel sensors and machine learning for predictive maintenance in medium voltage switchgear to enable the energy and mobility revolutions. *Sensors*, *20*(7), 2099.

Raue, A., Karlsson, J., Saccomani, M. P., Jirstrand, M., & Timmer, J. (2014). Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*, *30*(10), 1440–1448.

Stosur, M., Szewczyk, M., Sowa, K., Dawidowski, P., & Balcerek, P. (2016). Thermal behaviour analyses of gas-insulated switchgear compartment using thermal network method. *IET Generation, Transmission & Distribution*, *10*(12), 2833–2841.

BIOGRAPHIES

Giacomo Garegnani is a scientist at ABB corporate research. He obtained a PhD in Mathematics from EPFL in 2021, with a thesis on inverse problems involving partial and stochastic differential equations, and on probabilistic numerical methods. His research interests include uncertainty quantification of numerical solvers, model identifiability, and statistical inference for condition monitoring.

Kai Hencken is a corporate research fellow at ABB corporate research. He obtained a PhD in Theoretical Physics from the University of Basel in 1994. He was a post-doc at the University of Washington from 1995 to 1997 and at the University of Basel from 1997 to 2005, where he received his Habilitation in 2000 and is a lecturer since. In 2005 he joined the theoretical Physics group at ABB corpo-

rate research. His research interests include the combination of physical modeling with statistical methods to solve problems related to industrial devices, as well as developing diagnostics and prognostics approaches.

Frank Kassubek obtained a PhD in Physics from the University of Freiburg in 2000 (“Electrical and Mechanical Properties of Metallic Nanowires”). At ABB corporate research, he works on a wide range of topics including modeling of electrical systems and sensors, plasma and arc physics, and PHM topics.

A Practical Example of Applying Machine Learning to a Real Turbofan Engine Issue: NEOP

Zdenek Hrncir[1] and Chris Hickenbottom[2]

[1]*Honeywell International, Brno, Czech Republic*
zdenek.hrncir@honeywell.com

[2]*Honeywell International, Phoenix, AZ, USA*
chris.hickenbottom@honeywell.com

ABSTRACT

There are high expectations for the use of Machine Learning algorithms in Engine Health Management, but the practical application for use with turbofan engines is often hindered by small sample sizes and noisy data. This paper discusses a case in which Machine Learning techniques were combined with domain expertise to develop a classifier called Non-seal Erratic Oil Pressure (NEOP). This classifier is used as an engineering tool to support manual review of engines flagged with Honeywell's OPX (Oil Pressure Transducer) algorithm. The purpose of the classifier is to assist a human in analyzing engine trend data from the HTF7000 turbofan engine, when the OPX algorithm identifies an engine with erratic oil pressure. The NEOP history provides an additional data source when deciding if aft sump maintenance is needed to replace a worn carbon seal, or if the erratic signal is associated with some other cause. The OPX algorithm has enabled the prevention and avoidance of costly unscheduled engine failures resulting in millions of dollars in documented savings, and the NEOP algorithm helps to ensure that the conclusions from the OPX process continue to result in the appropriate engines being identified for maintenance inspection and corrective action.

1. INTRODUCTION

Data science and machine learning techniques hold great promise in the realm of proactive engine health monitoring, but currently there is a considerable gap between the conceptual possibilities and real-world results. This paper discusses an example where machine learning techniques, guided by domain expertise, were successfully utilized to produce an algorithm with real value.

Honeywell Aerospace manufactures the HTF7000 turbofan engine that powers several super-mid-size (SMS) business jets. Honeywell also develops Engine Health Monitoring algorithms to detect anomalies in the trends for those engines, indicating the presence of an incipient fault. These algorithms provide business jet operators with the ability to perform maintenance before the incipient fault progresses into a disruption to flight operations. A good example of these algorithms is the Carbon Seal Bimodality algorithm from OPX (Oil Pressure Transducer). Previous work (Hickenbottom, 2022) showed that this algorithm has proven very effective at detecting accelerated wear in the carbon seal near the number 4 bearing in the turbine section. It has correctly identified hundreds of engines with excessive carbon seal wear and allowed thousands of others to remain in service given evidence of healthy seals.

Once the Carbon Seal Bimodality algorithm and support process matured to the point that it can detect very small levels of seal wear, it became more prone to pick up other causes which present similar symptoms. After identifying a few false positive indications of carbon seal wear, a machine learning algorithm was developed to classify variability in oil pressure residual signature as either caused by Seal Wear, or Other Cause.

2. HISTORY OF CARBON SEAL BIMODALITY

The Carbon Seal Bimodality algorithm initially came into existence because of a need to detect incipient faults in the Oil Pressure Transducer (i.e., OPX). The first step was to correct the measured oil pressure because the measured pressure varies greatly with the oil temperature and engine operating regime. These normal variations can mask changes in oil pressure which are the symptoms of engine faults. The objective when developing the oil pressure correction logic was to use data science methods to analyze field data and identify the primary drivers of variation in the measured oil pressure. Once we determined the most 'correctable'

operating regime and we accounted for variations due to environmental conditions, we derived a model from field data. Comparing each oil pressure measurement to this model resulted in the Oil Pressure/Temperature Residual (OilPT Residual) CI, which is trended over time, with the initial intent of detecting incipient faults in the oil pressure transducer. By analyzing a handful of known OPX sensor failures we determined that a faulted sensor will often cause a shift or drift in OilPT Residual before the sensor fault progresses to the level detectable by the engine controller. Figure 1 is an example of the signature for a faulted oil pressure sensor.

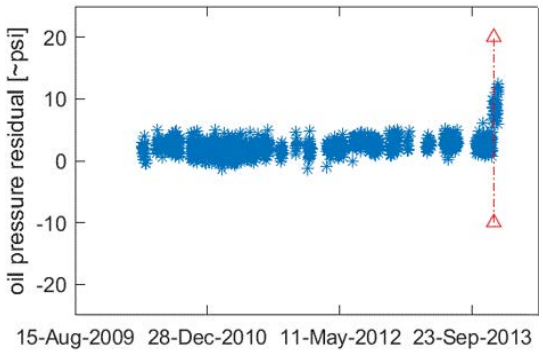


Figure 1. Trend of OilPT Residual with faulted oil pressure sensor

As we analyzed fleetwide trends of the OilPT Residual CI, we started to notice a unique pattern, where over time the CI would start to split into two separate populations, which would continue to diverge. Figure 2 is an example of an OilPT Residual trend with a bimodal distribution. The term bimodal refers to the two distinct peaks in the probability density function on the right side of the figure.

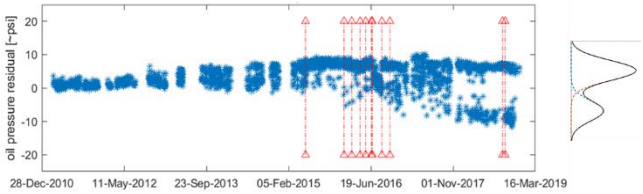


Figure 2. Bimodal distribution of OilPT Residual

As we searched for an explanation for this signature, we started thinking about the fact that the measured oil pressure is not an absolute measurement but is in fact a delta-pressure relative to the aft sump pressure. This implies that a perceived drop in oil supply pressure could be the result of an increase in the aft sump pressure. Next, we investigated the hypothesis that whatever might be causing the higher sump pressure would return to normal after hot section maintenance. To test this, we did a fleet run and identified several engines that had high bimodality at some point, which then went away abruptly. We then investigated the maintenance records for those engines and confirmed that the disappearance of bimodality correlated with the timing of hot section maintenance. This represented significant evidence to support the hypothesis that an increase in aft sump pressure is the cause of bimodality.

Once it became clear that hot section maintenance was causing the bimodality to reset, we started looking more closely at a carbon seal in the aft sump. An opportunity to inspect an engine with high bimodality presented itself and the condition of the carbon seal unlocked the mystery of OilPT Residual bimodality. Figure 3 shows the first carbon seal removed proactively based on bimodality in the OilPT Residual trend. Note that the pressure balance features seen on the top half of the picture on the right were originally present in the bottom half as well.

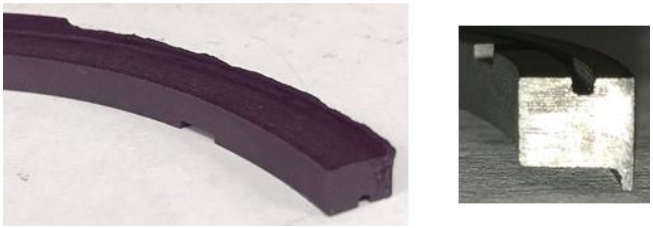


Figure 3. Worn carbon seal

With this new understanding of the correlation between bimodality and carbon seal wear, we conducted a fleet run and identified engines with varying degrees of carbon seal wear. As more engines with bimodality were inspected, the relationship between bimodality and carbon seal wear became even clearer. The strong correlation between bimodality and seal wear allowed the Service Related Difficulty investigation to focus on the engines with the highest level of wear, avoiding a fleetwide campaign of all fielded engines to replace the carbon seals with a new design. We began proactively removing carbon seals, which provided additional details to assess seal wear progression. Increased seal wear may result in secondary damage to the LP stub shaft (see Figure 4), which can increase maintenance costs. Being able to detect wear and replace the seal prior to this secondary damage results in significant maintenance cost savings. Since the operator can replace the seal while the engine is on the aircraft and wear occurs over hundreds of hours of operation, early detection also allows operators to address the issue without affecting their flight operations. The opportunistic maintenance from these alerts has resulted in millions of dollars in cost avoidance and improved aircraft uptime and availability.

Figure 4. Expensive secondary damage to stub shaft

3. NEED FOR NEOP ALGORITHM

At first, only those engines with the most severe seal wear were flagged to have their carbon seals replaced. As the improved-design carbon seals became more readily available, the bimodality threshold was gradually made more sensitive, such that more carbon seals were replaced earlier in their wear progression. This increased sensitivity means that variability in the data due to causes other than carbon seal wear can drive the bimodality measurement over the threshold.

Figure 5 is an example where an OilPT Residual trend is bimodal, but the bimodality is driven by a cause other than carbon seal wear. In this case, an alert was generated based on a very conservative assessment of the trend. Even though the review team felt it was unlikely that carbon seal wear was causing the bimodality on an engine with so few hours, the decision was made to enter the engine to inspect the carbon seal. This inspection revealed a healthy carbon seal, meaning that the alert was a False Positive.

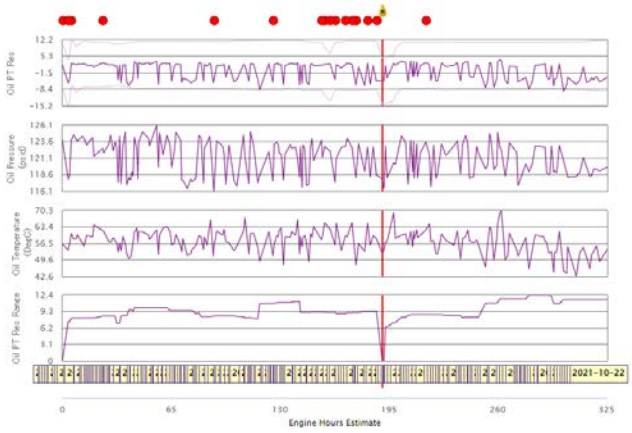


Figure 5. Bimodality driven by non-seal ("Other") cause

Prior to this case, the carbon seal bimodality algorithm had not resulted in any False Positive alerts to the aircraft operators. There were other examples of OilPT Residual trends with high variability, but they were visually determined to not fit the signature of carbon seal wear. Figure 6 is an example of a trend which was flagged by the algorithm, but manually overridden based on visual review by a domain expert.

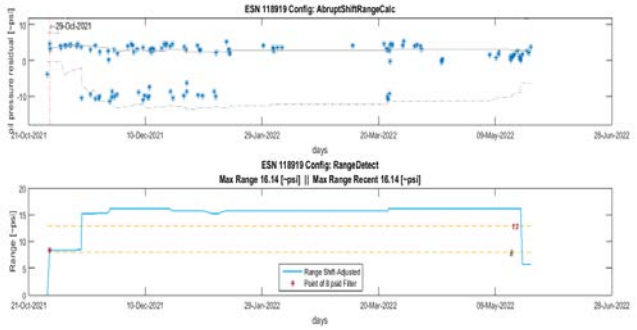


Figure 6. Expert determined bimodality not driven by seal wear

In some cases, it was easy for the review team to conclude that the variability in OilPT Residual was not caused by carbon seal wear, but in other cases it was not as clear. Figure 7 shows an example where visual review of the data did not result in an obvious conclusion. Because of the earlier False Positive, and the increasing number of cases where visual review of the data did not reveal an obvious conclusion, the team began investigating if a Machine Learning algorithm could be trained to distinguish carbon seal wear from other causes of variability in the OilPT Residual trend. This algorithm became known as NEOP (Non-seal Erratic Oil Pressure).

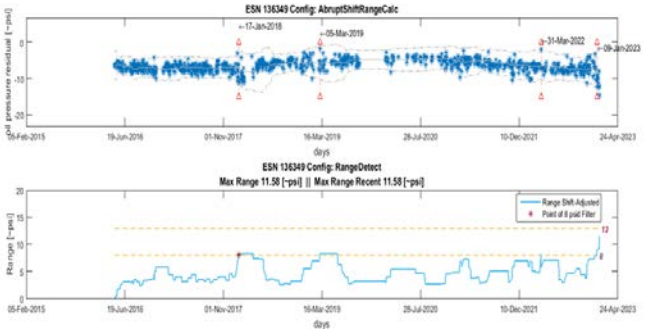


Figure 7. Cause of bimodality not obvious

4. ALGORITHM STRUCTURE

The NEOP algorithm is based on iterative development that progressed along with increased knowledge about collected oil system data and demand for further explanation of observed deviations from the model available at that time. The simplified diagram shown in Figure 8 involves the following steps:

• Data Filtering: this step applies known oil measuring system design limits to filter out invalid data, fuse data from multiple sources and in general assure that time series pressure and temperature data are of high quality.

• Oil Pressure Correction: applies known design factors that contribute to variability in measured Oil Pressure. These

corrections are not driven by data, they were engineered based on domain knowledge.

• Oil P/T Curve Residual: applies simple regression model that was trained from data across the fleet. The model captures the relation between oil pressure and temperature. This step eliminates the effect of oil viscosity on the flow of oil through the system and sensed oil pressure. Oil temperature is the data source that influences viscosity and can be smoothly correlated to oil pressure.

• Shift Adjustment Logic: applies detection of sudden shifts in Oil Pressure Residual to determine if there was a maintenance action to adjust oil pressure. This logic then eliminates the effect of the maintenance action to allow proper assessment of bimodality.

• Bimodality Detection Logic: OilPT Residual range proved to be good indicator of carbon seal wear.

• Calculate NEOP Features: extracts features for Non-seal Erratic Oil Pressure detection. Features are discussed in detail in section 7.

• Predict NEOP Class scores: Support Vector Machine classifier was trained and applied. One of its benefits over other classification techniques available in the legacy development environment in use (MATLAB 2015) is its ability to produce class scores, or confidence. These class scores are used to plot a continuous trend of the classification result, which is more informative than a binary output from decision trees, for example.

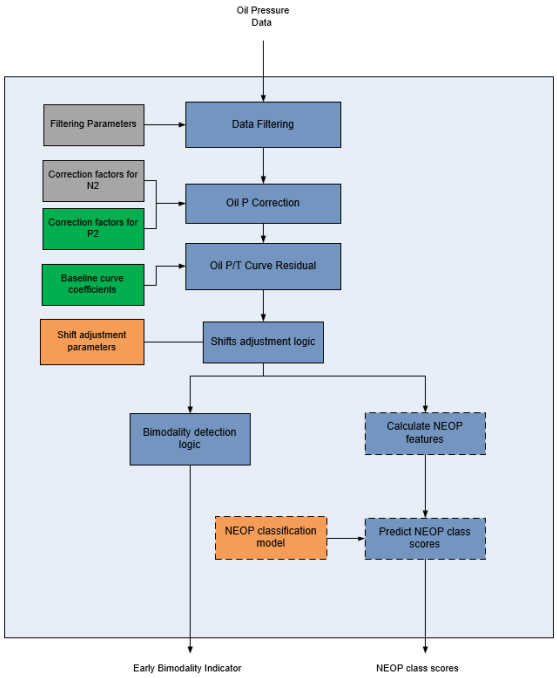


Figure 8. Structure of NEOP algorithm

5. TRAINING DATA

Aircraft engines are relatively low volume and high reliability assets. As a result, a very common problem when developing a diagnostic algorithm using a machine learning approach is shortage of training data for the fault cases. There is a huge imbalance between healthy and fault data. For the NEOP classifier training there were only a handful of engines exhibiting bimodality that, based on ground truth, could not be attributed to carbon seal wear. We'll denote these cases as "O" or Other Cause of bimodality. To describe the data, we used these groups:

• Healthy data: OilPT Residual with smooth trendline

• Severe carbon seal wear: OilPT Residual with very large bimodality in trendline

• "S" - Seal wear: OilPT Residual trend before the carbon seal replacement exhibiting the pattern of medium wear of the seal. See Figure 9.

• "O" - Other cause of bimodality (non-seal erratic oil pressure): OilPT Residual trend with known healthy seal but showing bimodal behavior that would be detected by Bimodality Detection logic and (incorrectly) marked as medium seal wear. See Figure 10.

Note that Figure 2 shows the characteristic progression through different data groups: from healthy data through "S" (Seal wear) to severe carbon seal wear.

The goal of this setup was to narrow down the classification problem to either class "S" or "O". This classification is only necessary during a portion of the fault progression. In early phases of wear, the bimodality range is low, and the original algorithm will correctly decide not to flag the engine for maintenance. For the advanced phases of wear, the fault signature changes, which would require the classification technique to learn a different pattern. Since the review team can visually classify advanced wear due to the signature over time, we decided to make a simplifying decision to focus the NEOP algorithm only on the middle phase of wear progression. As a result, severe carbon seal wear was excluded from the "S" group.

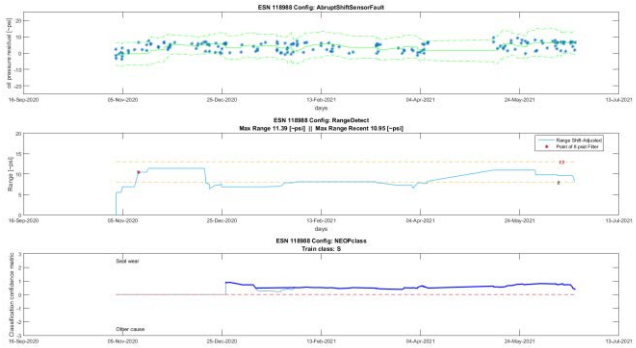


Figure 9. Training data example: "S" – Seal wear

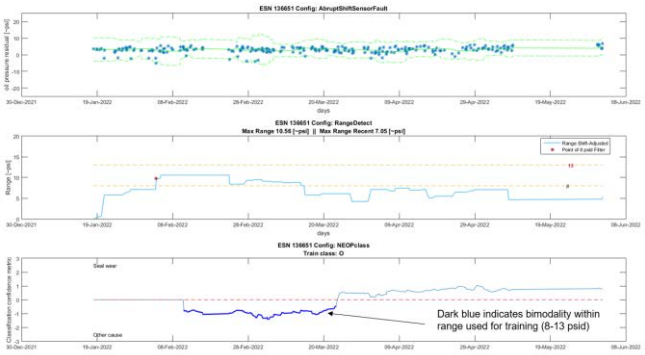


Figure 1010. Training data example: "O" - Other cause

Each engine that was included in the dataset provided one or more data series belonging to one of the groups. This led to variable lengths of OilPT Residual data series. To obtain a reasonable number of training samples we took several windows of 50 datapoints from each series. These windows were partially overlapping. The window size and overlapping step was chosen carefully to balance the need of having enough training samples and the need to have those samples be reasonably independent.

6. Configuration Management of Training Datasets

As discussed in the Society of Automotive Engineers Aerospace Information Report AIR6988, *Artificial Intelligence in Aeronautical Systems: Statement of Concerns*, one of the key considerations for developing and maturing a Machine Learning algorithm in an application like this is configuration management of training datasets. While configuration management and versioning of software modules is a well-understood activity in aerospace, configuration management and versioning of training and validation datasets used for machine learning is not as mature.

To ensure that the ML results were reproducible, and to enable iterative improvements as new cases became available for training, a repository was set up to store and version-control datasets. Standard naming conventions and processes were established so that multiple software developers could access the datasets and replicate each other's results.

7. Features

Features are calculated for the moving window, which moves along the timeline. The size of the window was set to be consistent with the bimodality detection logic: samples from 50 consecutive take-offs. This window is large enough to account for the fact that the bimodality signature in data was seen to temporarily cease for many consecutive datapoints.

The following features based on OilPT Residual were included in the final set:

• Range: this simple feature assures consistency with the previously implemented seal wear bimodality detector.

• Sigma (standard deviation): supplement to range feature.

• Gaussianity (fitness to gaussian distribution): this is the key measure that helps distinguish between noisy unimodal data and bimodal distribution.

• Skewness: it was observed that when bimodality starts occurring, the "S" class appears to have more evenly distributed datapoints between high and low OilPT Residual populations (skewness close to zero). While "O" class samples appear to have more occasional drops in OilPT Residual (negative skewness).

• Scatteredness: none of the measures listed above considers the order of datapoints inside the window. Although scatteredness is not a formally defined statistical measure, it is what we call what was implemented as RMS (Root Mean Square) of differences between consecutive points. This measure gives high values when OilPT Residual values are alternating between low and high values. This behavior is expected in medium seal wear. Domain knowledge of how the seal physically behaves in the engine (a worn seal randomly settles in one of two extreme positions where it's sampled during takeoff) enabled us to engineer this custom feature.

These features calculated on "S" and "O" training datasets were used to train the final Support Vector Machine classifier with Gaussian (or Radial Basis Function - RBF) kernel. Hyperparameter Kernel Scale was used to prevent overfitting to the training data. By tuning the kernel scale, we intentionally trained a medium-to-coarse model (in MATLAB Classification Learner terms) for the price of slightly decreased accuracy of the learned classifier. This setting was chosen to compensate for the fact that training samples were not perfectly independent, because they were taken from a limited set of engines. This fine-tuning is one example of using engineering experience and evaluation of individual plot results with analysts, rather than pure optimization of a goal metric, which is common in ML tasks with an abundant and balanced set of training data.

8. Example Cases

To illustrate how the NEOP output is interpreted, 4 real cases are discussed here. The first example, shown in Figure 11, is a straightforward case where the NEOP output (shown as 'classification confidence metric' in the third data series) is consistently above zero, indicating that the level of bimodality (shown as 'Range' in the second data series) can be attributed to real seal wear. This is useful to the review team because it increases confidence that an engine flagged for seal wear will not result in a False Positive disruption.

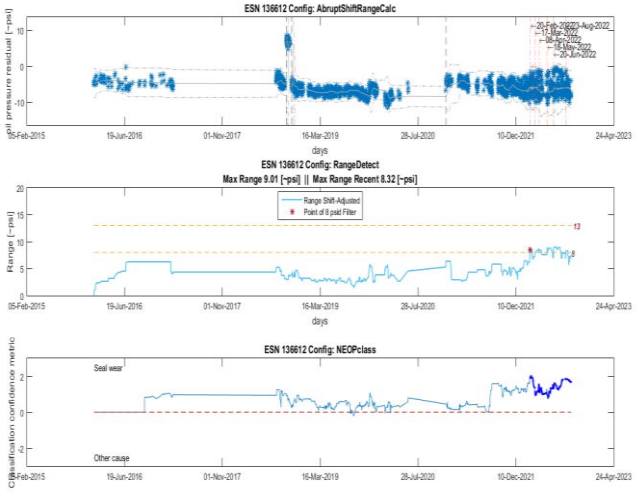


Figure 11. Consistently classified as seal wear

The second example, shown in Figure 12, illustrates how real-world limitations in data can affect the NEOP output. In this case there were large gaps in the data history. This caused the NEOP output to incorrectly interpret shifts as 'other cause', but the review team was able to use the NEOP output not affected by the data gaps to confirm that the carbon seal was worn. This is a good example where even when the ML algorithm encounters data outside of its training, an expert reviewer can still make sense of the data.

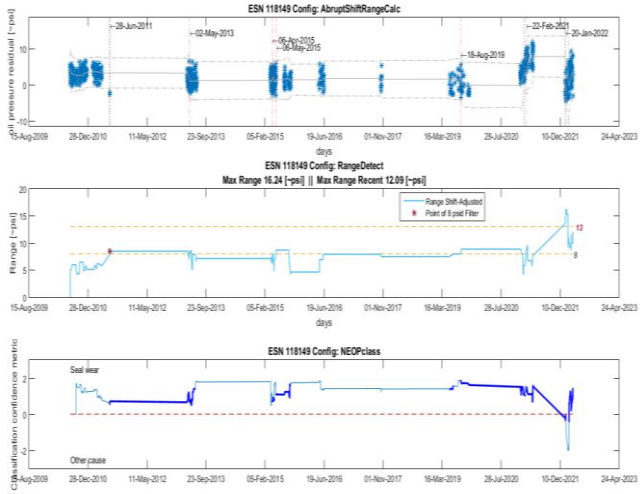


Figure 12. Large gaps in the data history

The third example, shown in Figure 13, is typical of the cases which motivated the creation of the NEOP algorithm. The bimodality range exceeds the threshold for seal wear, but the engine is known to have a healthy seal. The NEOP history in cases like this allows the review team to override the alert for carbon seal wear. Since there is no known operational impact associated with the 'other cause' classification, no supplemental maintenance is recommended.

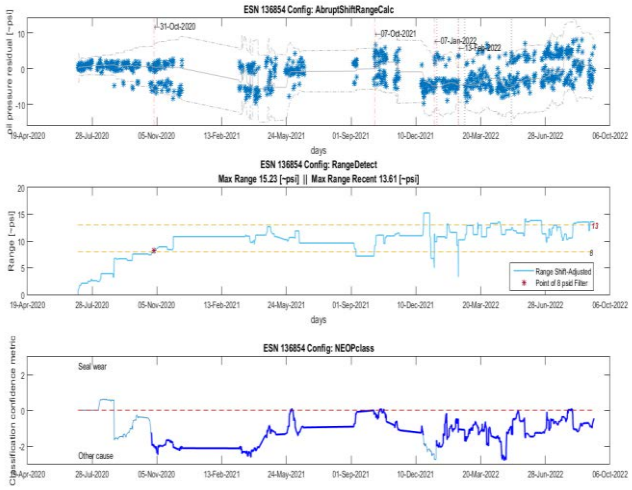


Figure 13. Healthy seal case with correct classification

The fourth example, shown in Figure 14Figure 13, is a case where the NEOP output moves back and forth between seal wear and 'other cause'. This is because the outliers which drive the bimodality range are intermittent, with periods of normal seal wear in between. By looking at the NEOP history, the review team can determine the true level of seal wear, and override the alert driven by the 'other cause' outliers.

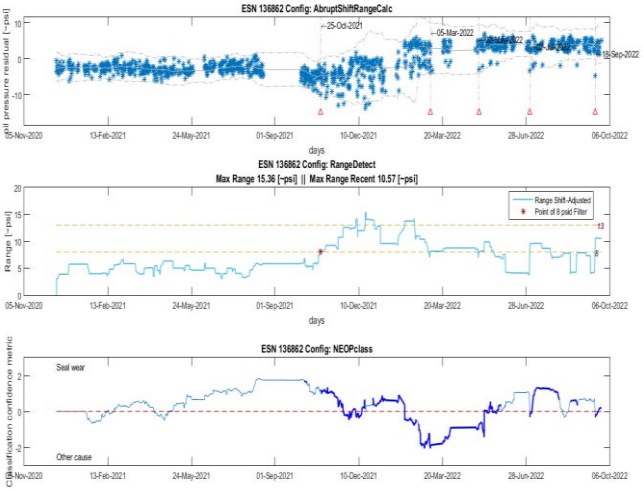


Figure 14. Intermittent outliers causing alternating classification

9. Combining Machine Learning with Domain Expertise

One of the fundamental lessons we've learned is that in applications like health monitoring of turbofan engines, synergy can be achieved when data scientists work closely with domain experts. Figure 1515 shows how these two groups of people make each other better. Data scientists are often able to use Machine Learning to identify relationships (correlations) between data. Domain experts can usually help the data scientist understand which correlations are meaningful (i.e., identify causation), and which correlations are trivial or meaningless. In doing this, the domain expert often learns more about their system, which in turn enables them to provide improved guidance for the next round of data science or machine learning.

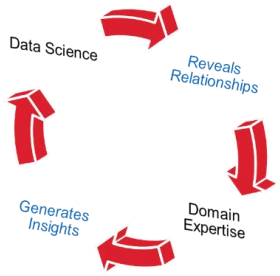


Figure 1515. Synergy between Data Science and Domain Expertise

In the case of the NEOP algorithm, several decisions were made by a domain expert to simplify the problem statement. For example, rather than requiring the algorithm to output a single answer, we recognized that showing the time-history and allowing a person to make a judgement is sufficient for the review team to decide on whether to flag an engine for seal wear. Another example is how the training data was limited to the time when OilPT Residual is between 8 and 13 psid. The data scientist learned that the algorithm did not train well across all OilPT Residual ranges. The domain expert recognized that there is a particular band of ranges where the interpretation is most critical, and the data scientist was able to refine the algorithm to focus on this area. The outputs from this refinement then helped the domain expert understand what is physically happening on the engine in these areas.

There are many examples where this synergy results in the data scientist making the domain expert more informed, and the domain expert contributing to making the data science more effective, which then provides additional information and feeds the cycle. The key is to have interactions early and often between the data scientist and the domain expert. This has proven to be much more effective than either a domain-independent data science approach or a purely expert-driven approach. For NEOP, this has resulted in the review team reviewing NEOP results 1-2 times per week, with the NEOP outputs being the key factor in the decision of whether to enter the engine in roughly 90% of those cases. Without this algorithm, many of those cases could result in unnecessary maintenance or failure of a carbon seal in flight.

10. Conclusion

As can be seen from the examples above, the NEOP output requires expert interpretation. Even though the algorithm does not provide a precise classification 100% of the time, it does provide valuable information which is of a practical benefit to the review teams. Often it is the simplifying assumptions/decisions like this which can move a potential machine learning approach from a great concept to a usable algorithm. With time, additional algorithm training could improve the ability of the NEOP algorithm to consistently classify the cause of wear, with less dependency on a domain expert; but even without improvement, the current algorithm has proven very valuable when the review team is faced with a signature that is difficult to explain.

References

Hickenbottom, C. D. (2022). Proactive approaches for Engine Health Management and a high value example. *Proceedings of IEEE Aerospace Conference*. 5-12 March 2022, Big Sky, MO. DOI: 10.1109/AERO53065.2022.9843255

Society of Automotive Engineers (SAE) (2021). Aerospace Information Report AIR6988, *Artificial Intelligence in Aeronautical Systems: Statement of Concerns*

Biographies

Zdenek Hrncir is a lead software engineer in Honeywell's Aero Analytics division. He received his master's degree in computer science from Masaryk university in Brno, Czech Republic in 2005. He dedicated most of his career to development of off-board diagnostic systems and analytics of Honeywell turbofan engines. He has also supported the Real-time on-board diagnostic system and flight controls of the Boeing 787. Zdenek is enthusiastic about ML and its application to aerospace machinery diagnostics.

A Review of Prognostics and Health Management in Wind Turbine Components

Jokin Cuesta[1], Urko Leturiondo[2], Yolanda Vidal[3], Francesc Pozo[4]

[1,2] *Artificial Intelligence and Data Area. Ikerlan Technology Research Center, Basque Research and Technology Alliance (BRTA). Pº J. Mª. Arizmendiarrieta, 2. 20500 Arrasate/Mondragón, Spain*
jcuesta@ikerlan.es
uleturiondo@ikerlan.es

[1,3,4] *Control, Data and Artificial Intelligence (CoDAlab), Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Campus Diagonal-Besòs (CDB), Universitat Politècnica de Catalunya (UPC), Eduard Maristany 16, 08019 Barcelona, Spain*
yolanda.vidal@upc.edu
francesc.pozo@upc.edu

[3,4] *Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya (UPC), Pau Gargallo 14, 08028 Barcelona, Spain*

ABSTRACT

Wind turbines (WTs) play an essential role in renewable energy generation, and ensuring their reliable operation is essential for sustainable energy production and reduction of levelized cost of energy. In this context, the field of prognostics and health management (PHM) is a powerful tool to predict and assess the health status of WT components, thereby enabling timely maintenance and reducing downtime. The study begins with an overview of WT components studied, including the blades, gearbox, generator, and bearings, and their common failure modes. For each component, various remaining useful life (RUL) estimation methods are explored, categorizing them into physics-based, data-driven, and hybrid methods. Despite the potential benefits, the application of PHM strategies in WTs is currently limited. Although PHM strategies have been present for years, their development in WTs remains a challenge. These key challenges are presented, including uncertainty management, integrating physical knowledge into models, variable operational conditions, data issues and system complexity.

1. INTRODUCTION

To meet the European Commission's target of achieving climate neutrality by 2050, reducing the levelized cost of energy (LCOE) is vital. According to the International Renewable Energy Agency (IREA, 2023), operation and maintenance (O&M) costs, which include fixed and variable components, typically constitute between 10% and 30% of the LCOE for most wind industry projects as of 2022. This underscores the importance of optimizing maintenance activities for wind turbines (WTs). It involves the transition from traditional corrective and preventive maintenance approaches to predictive maintenance strategies, where maintenance tasks are scheduled based on the real-time and projected condition of components. In this context, the field of prognostics and health management (PHM), which covers various techniques to monitor the evolution of component wear, plays a critical role. Through PHM, it becomes possible to forecast remaining useful life (RUL) of components using historical and current operational data (Ferreira & Gonçalves, 2022).

To effectively implement PHM strategies for WTs, it is essential to explore their target components and their failure modes. The main components of a WT's drivetrain include the rotor, gearbox, and generator, which are interconnected via the low-speed shaft (LSS) and high-speed shaft (HSS) (see Figure 1). The gearbox, generator, and rotor blades are identified as the most critical subsystems, both onshore and offshore, based on downtime analysis (Dao, Kazemtabrizi, & Crabtree, 2019). Moreover, failures associated with the gearbox, rotor blades, and generator represent a higher expenditure, in that specified order (Tazi, Châtelet, & Bouzidi, 2017).

The main failure modes that affect these components are the

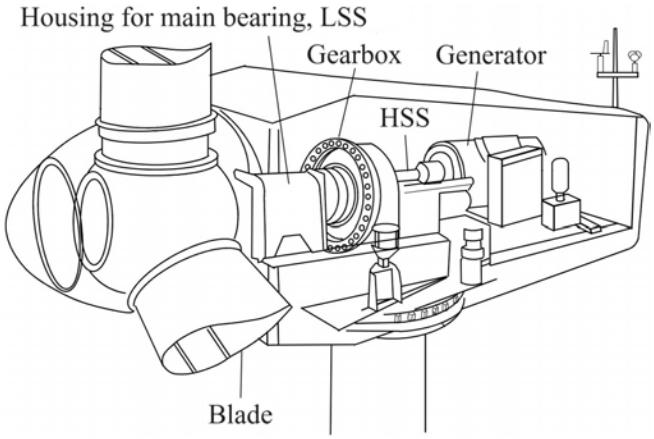


Figure 1. Schematic diagram of the components explored in this paper. Readapted from (Jiang et al., 2017).

following:

- **Blades**: fatigue, corrosion, and aerodynamic imbalance/ asymmetry (Catelani, Ciani, Galar, & Patrizi, 2020).
- **Gearbox**: abrasive wear, pitting, cracking, scuffing oil leakage, insufficient lubrication (Owolabi, Madushele, Adedeji, & Olatunji, 2023; Olabi et al., 2021).
- **Generator**: overspeed, overheat, wear, excessive vibration, rotor asymmetries, bar break, electrical problems (Olabi et al., 2021; Lydia & Edwin Prem Kumar, 2023).
- **Bearings**: axial cracking, spalling, pitting, brinelling (fretting) (Owolabi et al., 2023).

Based on the needs for PHM implentation in WT components, the scope of this review paper is to provide an in-depth analysis of the methodologies, algorithms, and techniques used to estimate the RUL of components within WT components. The aim of this review is to present works published from 2018 to March 25, 2024, thereby gathering recent advancements and trends in PHM specific to critical wind components, including blades, gearboxes, bearings, and generators. By addressing these challenges and providing a comprehensive review of the current state-of-the-art, this paper aims to contribute to a better understanding of the complexities and future research trends involved in developing RUL prognostics for wind turbines. The paper is structured as follows. Section 2 consolidates the research efforts made in the prediction of RUL classified by the component to which the techniques are applied; finally, Section 3 focuses on conclusions and key challenges.

2. Prognostics. RUL estimation

Prognostics refers to the examination of fault symptoms to forecast future conditions and RUL within designed parameters (ISO, 2012). This section aims to gather the works done for accurate prediction of RUL in WT components, classifying the methods into physics-based, data-driven and hybrid. The applications of components found coincide with the most critical components in terms of dowtime and repair costs mentioned above, classified in blades, gearbox, generator, other bearings (which include predictions of RUL of bearings whose location is not specified) and those that consider WT as a system. It is important to note that most of the works found focus on bearing prediction, many of them located on the HSS. These can be gearbox high-speed bearings, gearbox intermediate-speed bearings, and generator bearings (Z. Liu & Zhang, 2020). When the paper introduced specifies the location of these, they are included in the gearbox/generator subsection. If not, they are included in other bearings.

The distribution of eighty-one papers among years and components can be found in Figure 2. It can be seen that data-driven approaches are the most common ones to predict the RUL of WT components, and there is an increasing trend towards using hybrid models (Figure 2a). Furthermore, the components most studied have been the gearbox and the generator, respectively (Figure 2b). Figure 3 gathers all the methods found in the literature, classified by type and component.

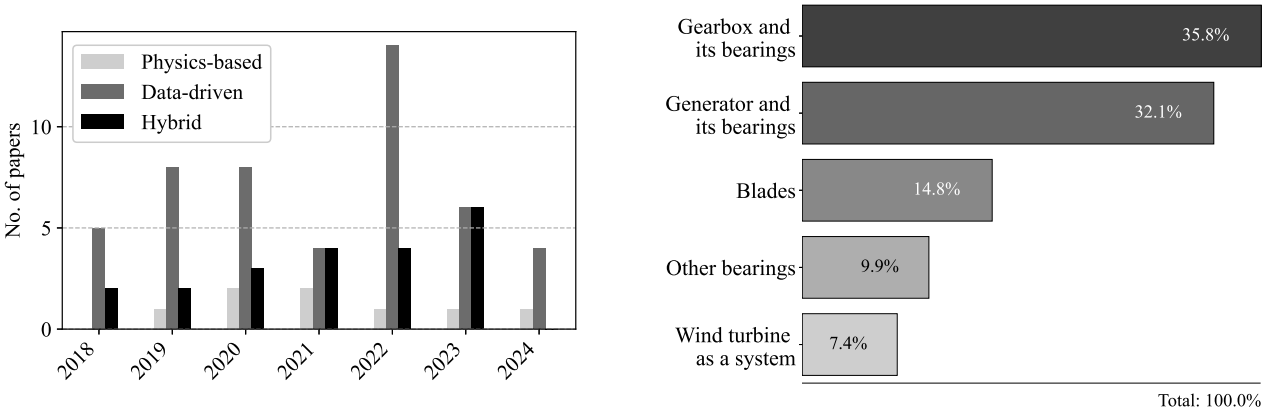
2.1. Blades

WT blades are engineered for a minimum of 20-year lifespan, resulting in load cycles between 10 millions and one billion, making them very susceptible to fatigue (Moroney & Verma, 2023). Twelve studies have been found to estimate the RUL of WT blades.

2.1.1. Physics-based models

Physics-based models have been used for WT blades RUL estimation. Studies such as (Saathoff, Rosemeier, Kleinselbeck, & Rathmann, 2021) and (Moroney & Verma, 2023) employed aeroelastic load simulations, and durability and damage tolerance analysis (DADTA), respectively, to quantify the effects of factors like blade pitch misalignment, and material fatigue on RUL.

Furthermore, one of the most widely used physics-based techniques has been Kalman filtering. (Muto, Namura, Ukei, & Takeda, 2019) proposed a method that combines load monitoring with dynamic response estimation, enhancing the accuracy of RUL evaluation with Kalman filter (KF). (Boutrous, Puig, & Nejjari, 2022) introduced an innovative model-based prognostics procedure, leveraging zonotopic KFs to quantify uncertainties in degradation propagation. Moreover, (Vettori, Lorenzo, Peeters, Luczak, & Chatzi, 2023) presented an adaptive noise augmented KF approach, addressing challenges in noise calibration for joint input-state estimation. Their method demonstrated superior performance in virtual sensing (VS) applications in diverse structural scenarios.



(a) Number of papers found in this review, categorized by modeling approach (physics-based, data-driven, and hybrid), over the years.

(b) Percentage distribution of RUL prediction techniques found across WT components.

Figure 2. Classification of papers in this review: a) by year and type b) by component.

2.1.2. Data-driven models

Among data-driven methods that have been used to estimate the RUL of WT blades, particle filter (PF)-based approaches offer a dynamic and versatile solution. Studies conducted by (Valeti & Pakzad, 2018, 2019), (Jaramillo, Gutiérrez, Orchard, Guarini, & Astroza, 2022), and (Lee, Roh, & Park, 2022) demonstrated the efficacy of PF in accurately predicting RUL of blades under varying conditions, including fatigue damage and dynamic loading scenarios. Alternatively, various methodologies that employ data-driven strategies, particularly those using artificial intelligence (AI), presented different avenues. For instance, the work introduced by (Yue, Ping, & Lanxin, 2018), an end-to-end model based on convolutional neural network (CNN) combined with long short-term memory (LSTM) networks, exemplifies such approaches.

2.1.3. Hybrid models

Hybrid models offer promising results for an accurate prediction of the RUL of WT blades. (Rezamand et al., 2021a) introduced an integrated fuzzy-based failure prognosis method, leveraging recursive principal component analysis (PCA), a wavelet-based probability density function (PDF) estimation, a Takagi-Sugeno (T-S) fuzzy system, and a Bayesian algorithm. Their approach enabled real-time predictions by capturing blade failure dynamics, categorizing nonlinear degradation trends, and estimating RUL for each trend, culminating in an aggregated prediction for the entire system. Applied to supervisory control and data acquisition (SCADA) data from real wind farms, the methodology demonstrated robust performance, outperforming traditional Bayesian methods and effectively modeling nonlinear failure dynamics. In another study, (C. Peng, Chen, Zhou, Wang, & Tang, 2020) focused on improving the accuracy of icing failure prediction in WT blades through a novel balancing algorithm based on boundary division synthetic minority oversampling technology (BD-SMOTE) and a multi-step prediction process using multiple Elman neural networks (ENNs).

2.2. Gearbox

Gearboxes operate under harsh environmental conditions, including vibrations from turbine-side components and wind, as well as fluctuations from the load through the generator, while stepping up the speed from the LSS to meet the requirements of the HSS that drives the generator (Salameh, Cauet, Etien, Sakout, & Rambault, 2018). Their failures contribute to around 20% of WT downtime (Lydia & Edwin Prem Kumar, 2023); therefore, it is essential to accurately predict their RUL. Twenty-nine works have been identified.

2.2.1. Physics-based models

In the field of WT gearbox reliability estimation, only one work has been found in the literature. (Pagitsch, Jacobs, & Bosse, 2020) presented a pioneering approach for modeling WT gearboxes with minimal parameters, emphasizing its utility in estimating RUL and facilitating real-time condition monitoring (CM). The determination of forces and bending moments acting on the main components involves using information on non-torque loads from the rotor sub-model and rotor torque from the SCADA data record. These inputs are then employed to calculate inner loads on machine elements in the gearbox, using rigid beam models and analytical basic equations for a three-stage WT gearbox intermediate-speed shaft bearing forces. Finally, the modified life rating as defined in ISO 281:2007 is applied to predict RUL.

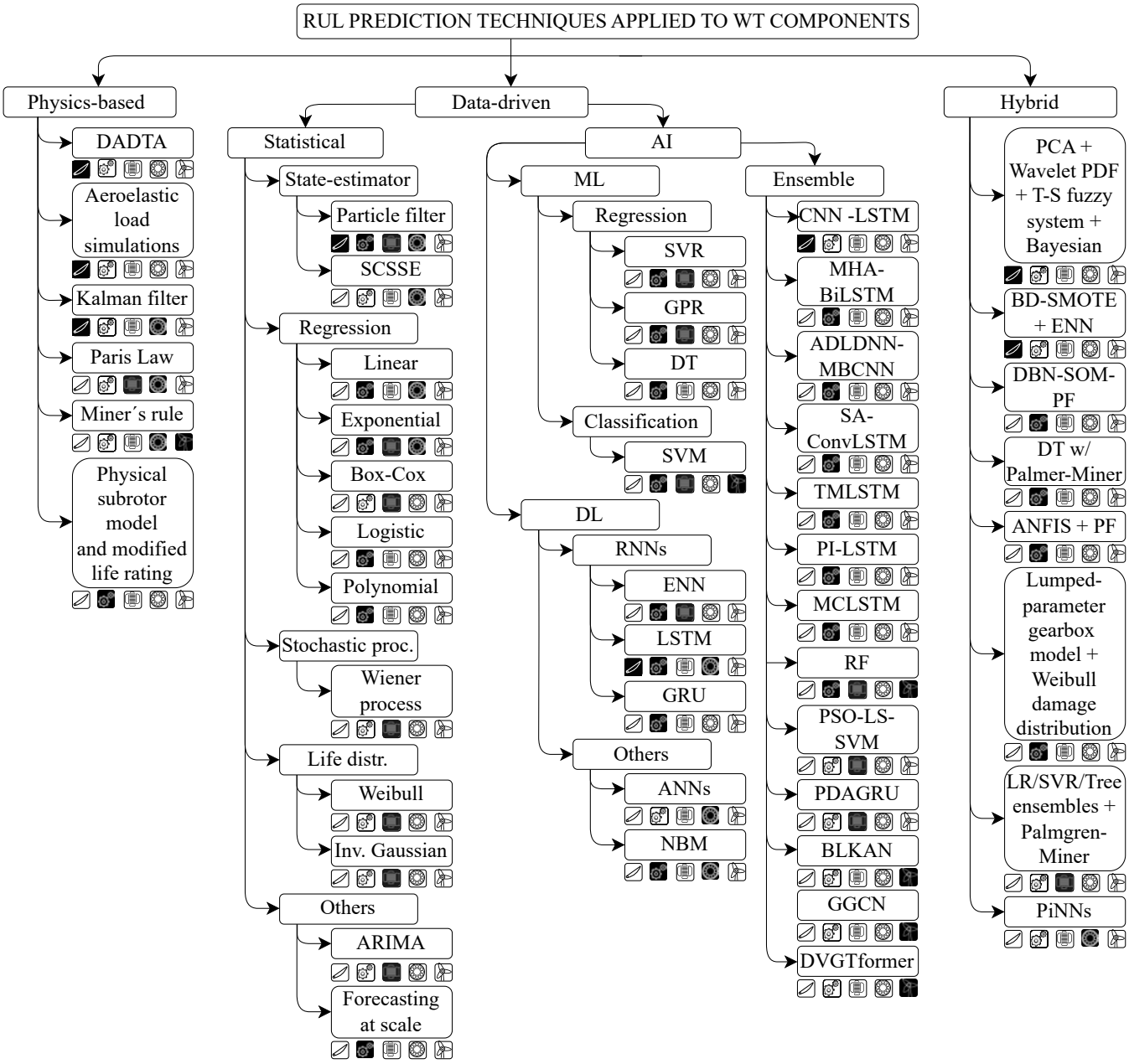


Figure 3. Techniques found in papers to predict RUL of WT components. The five categorical boxes correspond to, in order: blades, gearbox and its bearings, generator and its bearings, other bearings, and the WT as a system. The components highlighted with a black background represent those studied with the corresponding technique.

2.2.2. Data-driven models

Data-driven methodologies have gained attention to predict RUL of gearboxes. These approaches use advanced statistical methods, such as PF and regressions, and AI techniques such as advanced LSTM networks and artificial neural networks (ANNs).

Statistical PF methods have shown promising results. (J. Wang, Gao, Yuan, Fan, & Zhang, 2019) proposed an approach that integrates a PF with an expectation maximization algorithm, effectively predicting bearing defects from vibration signals in a 2MW WT gearbox. This method quantifies uncertainty in predictions, reduces false alarms, and highlights the importance of Bayesian inference for effective prognosis. (Cheng, Qu, Qiao, & Hao, 2019) introduced an enhanced particle filter (EPF) algorithm tailored for bearing RUL prediction in a 2.5 MW doubly fed induction generator (DFIG)-based WT gearbox. The EPF algorithm overcomes particle impoverishment

issues, demonstrating superior performance compared to traditional PF methods. Additionally, (J. Wang, Liang, Zheng, Gao, & Zhang, 2020) proposed a Bayesian framework integrating fault prognosis and PF-based RUL estimation, effectively predicting RUL while quantifying uncertainties.

The construction of a robust health indicator (HI) is crucial for effective RUL prediction. Methodologies often combine signal processing and statistical modeling to extract information from sensor data. Then, the trend of these indexes is estimated using different algorithms. For instance, (Praveen, Shah, Pandey, Vamsi, & Sabareesh, 2019) developed a HI from vibration signatures using wavelet transform and PCA, achieving high RUL prediction accuracy with an exponential degradation model. On the other hand, (Lázaro, Yürüen, & Melero, 2020) constructed a SCADA-based functional indicator (FI) methodology using Gaussian mixture copula model (GMCM) from SCADA signals.

Several studies have explored AI methods to predict RUL of gearboxes in WTs. Some of them have compared various techniques to determine the most effective approach. For instance, (Tayade, Patil, Phalle, Kazi, & Powar, 2019) explored regression models, such as support vector regression (SVR) and random forest (RF) regression, augmented with PCA for feature selection, demonstrating the superior accuracy of RF over SVR in early fault detection and performance degradation prediction. (Carroll et al., 2019) employed ANNs, support vector machines (SVMs), and logistic regression to predict failures, with ANNs outperforming other methods in accuracy, especially when using SCADA and vibration data. (Elasha, Shanbr, Li, & Mba, 2019) focused on gearbox bearing prognosis, demonstrating the superiority of exponential and polynomial regression models over multilayer ANNs, particularly in terms of root-mean-square error (RMSE) and R^2 coefficient. Lastly, (Elforjani, 2020) conducted a comprehensive comparison of machine learning (ML) techniques, revealing that Gaussian process (GP) exhibited the lowest error levels compared to decision trees, SVM and a feedforward ANN.

Traditional LSTM networks have widely been used for time series forecasting; nevertheless, they show some limitations in RUL prediction. Several recent works have aimed to address these issues, recognizing the challenges posed by their inability to effectively capture global trends over time and tap into backward and forward connections within time series data. (Shen, Tang, Li, Tan, & Wu, 2022) introduced the multi-head attention bidirectional-long-short-term-memory (MHA-BiLSTM), which incorporates a multi-head attention mechanism to dynamically weigh circulating data between cells, thereby enhancing the network's ability to focus on information crucial to the degradation process. In another study, (Xiang, Qin, Liu, & Gryllias, 2022) proposed the automatic multi-differential learning deep neural network (ADLDNN), leveraging a measurement level division unit and a multi-branch convolutional neural network (MBCNN) to address the varying input contributions over time, demonstrating superior performance over existing methods. (Xiang, Qin, Luo, & Pu, 2022) proposed the spatio-temporally multidifferential network (SMDN), which used temporally multidifferential LSTM (TMLSTM) and spatially multidifferential CNN (SMCNN) sub-networks to capture spatio-temporal information effectively, achieving superior performance in RUL prediction. Furthermore, (Xiang, Qin, Luo, Wu, & Gryllias, 2023) presented the concise self-adapting deep learning network (CSDLN), which integrates a multi-branch 1D involution neural network (MINN) and a multi-head graph recurrent unit (GRU) to dynamically extract hidden features and adaptively learn them, resulting in enhanced RUL prediction accuracy. In addition, (B. Li, Tang, Deng, & Zhao, 2021) introduced the self-attention ConvLSTM (SA-ConvLSTM), which combines ConvLSTM architecture with a self-attention mechanism to selectively focus on important information and improve training efficiency and prediction accuracy. Moreover, (Z. Wang, Gao, & Chu, 2022) presented the pre-interaction LSTM, designed to enhance the capture of sequential features in time-series limited samples, especially during periods of interrupted continuous feature. Lastly, (Xiang, Li, Luo, & Qin, 2024) introduced the multi-cellular long short-term memory (MCLSTM) to obtain distinct distributions of monitoring data and utilized domain adversarial and active screen mechanisms for transfer learning.

Efforts have been made to select suitable features and construct a HI to enhance RUL estimation with AI. Qin et al. introduced the shape-characteristic similarity autoencoder (SM-SAE) network to automatically extract HI curves with specific shape characteristics from raw sensing data, thereby improving degradation trajectory characterization (Qin, Yang, Zhou, Pu, & Mao, 2023). Similarly, He et al. present the self-calibration temporal convolutional network (SCTCN) model, leveraging multidomain feature extraction and a self calibration module for improved prediction accuracy, even with limited time series data (He, Su, Tian, Yu, & Luo, 2022).

One of the important data sources for predictive maintenance in WTs is SCADA data. (Verma, Zappalá, Sheng, & Watson, 2022) extensively explored the use of high-frequency SCADA data, employing advanced techniques to address imbalanced operational regimes and enhance detection capabilities in WT gearbox failure prediction, and using ANN-based normal behaviour model (NBM) and one-class SVM. In contrast, (Bermúdez, Ortiz-Holguin, Tutivén, Vidal, & Benalcázar-Parra, 2022) present an ensemble neural network model, combining a two-dimensional CNN for spatial information extraction and an LSTM network for spatio-temporal feature analysis. The model was trained only on data from SCADA (Bermúdez et al., 2022).

2.2.3. Hybrid models

Gearboxes have been the components to which most hybrid models found in this work have been applied. Desai et al. demonstrated the potential of integrating bearing-specific data from physics-based models with conventional SCADA data to enhance bearing failure prognostics (Desai, Guo, Sheng, Phillips, & Williams, 2020), highlighting significant improvements in F1 score and AUC. However, they suggested further refinement by developing individual models for each bearing type. Similarly, Mehlan et al. presented a VS method designed for online load monitoring and subsequent RUL assessment of WT gearbox bearings within a digital twin (DT) framework (Mehlan, Nejad, & Gao, 2022). The virtual sensor integrates data from readily available sensors in the condition monitoring system (CMS) and SCADA system with a physics-based gearbox model, employing multiple state estimation methods for load estimation and the Palmgren-Miner model for RUL assessment.

(Pan, Hong, Chen, & Wu, 2020) proposed a novel hybrid methodology which integrated deep belief network (DBN), self organizing feature maps (SOMs) and PF, DBN-SOM-PF. Their approach showcased superior performance in accurately predicting degradation tendencies and reducing RUL uncertainty. Moreover, (Cheng, Qu, & Qiao, 2018) introduced an adaptive neuro-fuzzy inference system (ANFIS)-based PF, demonstrating its superiority over traditional recurrent neural networks (RNNs). Their study addressed challenges in varying speed conditions through signal resampling, enhancing fault diagnosis effectiveness. (Qiao & Qu, 2018) also employed an ANFIS model in fault prognosis, showcasing accurate trend prediction validated by a gearbox run-to-failure test. Moreover, (Z. Li, Zhang, Kari, & Hu, 2021) proposed a comprehensive evaluation function combined with SOM network to construct a HI curve for gearbox-side high-speed shaft bearings (HSSB), then a Bayesian update model and expectation maximization algorithm were employed for RUL estimation. The model demonstrated superior accuracy in RUL prediction compared to SVR. Lastly, (Zheng et al., 2024) presented a multi-stage RUL prediction model tailored for WT planetary gearboxes, emphasizing interpretability and achieving promising results in real-world scenarios.

Finally, (Guo et al., 2020) integrated physics-domain models, SCADA data, and wind plant failure records to forecast the probability of failure for individual gearbox bearings. Focusing on bearing axial cracking, the study considers frictional energy accumulation and electrical power generation as prognostic metrics. The lumped-parameter gearbox model calculates gearbox bearing radial loads and displacements at any given torque, then Weibull distribution of the accumulated damage threshold of the accumulated energy is determined statistically.

2.3. Generator

Generators are labeled as critical components, as the O&M costs caused by the premature failure of the main components of WT generators can represent a significant portion –around 10-20%– of the overall energy expenses for a WT project (Cao et al., 2018). Twenty-six studies have been identified to predict the RUL of this component.

2.3.1. Physics-based models

Within physics-based models to predict the RUL of WT generators, Kalman smoother (KS) method has been uniquely identified. (Saidi, Ali, Benbouzid, & Bechhofer, 2018) introduced an integrated prognostics methodology for WT HSSB prognosis, focusing on bearing failure prognosis driven by excessive shaft vibration. Their approach used a usage model based on Paris' law and a KS to estimate RUL, addressing inherent phase delay cancelation from Kalman filtering for improved accuracy and smoother estimated with confidence bounds.

2.3.2. Data-driven models

In statistical methodologies, in both studies carried out by (P. Wang, Long, & Wang, 2020) and (Farhat, Chaari, Chiementin, Bolaers, & Haddar, 2022), the prediction of RUL for WT generator bearings is accomplished primarily through the implementation of exponential degradation models. Wang et al. used a fusion method based on PCA to construct a HI, which serves as a crucial metric reflecting the degradation level. This HI, alongside features extracted from vibration data and statistical analyzes such as monotonicity analysis and hierarchical clustering, contributes to accurate RUL estimation. Similarly, Farhat et al. also used an exponential degradation model for RUL prediction, initializing parameters based on healthy data and iteratively updating them as degradation progresses, obtaining a dynamic HI selection with good accuracy.

In their research, Rezamand et al. focused on reliability metrics for WT generators and real-time RUL prediction for critical bearings. In their study from 2019, (Rezamand, Carriveau, Ting, Davison, & Davis, 2019) explored reliability metrics using truncated WT generator data from a 100 MW wind farm, employing Weibull and accelerated life testing analysis to identify best-fitted distribution models and propose predictive PDF and hazard functions for the generator group. Subsequently, in their 2020 study, (Rezamand, Kordestani, Carriveau, Ting, & Saif, 2020) introduced a novel real-time Bayesian RUL prediction algorithm, incorporating comprehensive feature extraction, selection, and signal denoising techniques. It demonstrated superior performance over single-feature-driven Bayesian algorithms through experimental case studies, offering an improved approach to provide accurate RUL predictions by combining information

from various single features using an ordered weighted averaging (OWA) operator.

The use of the Wiener process to predict RUL is apparent in several studies focusing on WT bearing health prognosis. (Hu et al., 2018) proposed an RUL prediction model based on the Wiener process and inverse Gaussian distribution, specifically targeting rear bearings of a 1.5 MW WT generator. By establishing an inverse Gaussian distribution function and deriving drift and diffusion parameters, the model effectively predicts RUL based on temperature monitoring data, offering valuable insights for maintenance decision-making. In a similar approach, (M. Liu, Dong, & Shi, 2023) addressed challenges associated with traditional vibration data by introducing a nonlinear Wiener degradation model integrated with physical and data knowledge. Their approach, which incorporates multi-sensor temperature data fusion and Bayesian analysis, demonstrated superior accuracy and reliability compared to conventional models. Furthermore, (Song, Youliang, Kai, Cheng, & Tao, 2020) employed both linear and nonlinear Wiener processes to construct dynamic monitoring and performance degradation models, intricately linking bearing temperature parameters, wind speed, and time. Lastly, (Lan et al., 2023) developed a precise RUL prediction method for WT generator bearings using a nonlinear Wiener process. Their approach used the 3σ criterion for online monitoring, considering a two-stage evolution of bearing performance parameters.

Other statistical techniques have been used to improve RUL prediction and health state estimation in WT systems. (Y. Peng, Bi, & Wang, 2023) developed a model integrating an enhanced two-phase Box–Cox transformation into the switching state-space model, capturing nonlinear degradation with phase transition behavior. Their adaptive parameter learning method dynamically estimated transformation parameters, phase transition positions, and predicted uncertainty. In a different approach, (Peter, Zappalá, Schamboeck, & Watson, 2022) proposed a framework combining data preprocessing, anomaly detection, and time series forecasting using SCADA signals and one-class SVM. Time series is then forecasted using an autoregressive integrated moving average (ARIMA) mode. Additionally, (Kramti, Saidi, Ali, Sayadi, & Bechhoeferer, 2019) leveraged PF-based Bayesian inference with advanced signal processing methods like spectral kurtosis and high order statistics for health state estimation of the generator-side HSSBs, demonstrating promising results for RUL estimation.

On the use of AI, Cao et al. have contributed significantly to the field of WT generator bearing RUL prediction with a series of innovative approaches. In their 2018 work, (Cao et al., 2018) introduced the interval whitenization Gaussian process (IWGP) method, which integrates interval whitenization and Gaussian process algorithms to forecast RUL under non-stationary operating conditions. This method showcased notable improvements over SVR and ANN techniques. Building upon this foundation, their subsequent study proposed a more comprehensive methodology, incorporating empirical mode decomposition (EMD) for signal denoising and fault development featuress (FDFs) extraction, followed by SVR modeling (Cao, Qian, & Pei, 2019). Expanding further, their latest work introduced the parallel gated recurrent unit with dual-stage attention mechanism (PDAGRU) model coupled with a novel uncertainty quantification method, enhancing both prediction accuracy and uncertainty assessment (Cao, Zhang, Meng, & Wang, 2023). By integrating a dual-stage attention mechanism and employing kernel density estimation and Monte Carlo dropout, their approach achieved remarkable RUL prediction accuracy.

The use of neural networks (NNs) in prognostics of HSSB in WT generators has shown promising results in recent studies. (Kramti, Ben Ali, Saidi, Sayadi, & Bechhoefer, 2018) introduced an ENN architecture, employing statistical time-domain features extracted from vibration signals as inputs. Their model demonstrated reliable performance even in the presence of noisy measurements. Expanding on this work, (Kramti et al., 2021) proposed a novel feature selection method based on monotonicity, trendability, and prognosability metrics, enhancing the robustness of their ENN-based prognostic model. Similarly, (Merainani, Laddada, Bechhoefer, Chikh, & Benazzouz, 2022) developed an ENN-based approach, incorporating a novel HI derived from spectral shape factor entropy and the Teager energy operator. Furthermore, (Hayder & Saidi, 2021) proposed a deep learning (DL)-based approach using a multilayer NNs, emphasizing the significance of kurtosis as a HI.

Authors also have contributed ensembled AI models to improve the accuracy of predictions. (Pandit & Xie, 2023) introduced an innovative approach combining sparrow search algorithm (SSA) with SVM, RF regression, and Gaussian process regression (GPR). Their model, driven by vibration signal analysis and feature selection based on monotonicity, exhibited high performance. In contrast, (Du, Jia, Yu, Shi, & Gong, 2023) addressed the limitations of traditional CNN models in extracting critical features for RUL prediction of bearings. They proposed a CNN prediction model enriched with a global attention mechanism (GAM) to enhance prediction accuracy. By transforming one-dimensional vibration signals into two-dimensional image data suitable for CNN processing and incorporating a HI constructed based on time-domain degradation characteristics, their approach significantly improved RUL prediction performance.

(Dameshghi & Refan, 2021) presented an innovative framework for prognosis, focusing on the failure behavior of the DFIG due to rotor electrical asymmetries. Their approach integrates the CMS module with the prognosis module, em-

ploying agents like the degradation trend index to enhance RUL prediction accuracy. Using a particle swarm optimization (PSO)-least squares (LS)-SVM method, with parameter tuning for LS-SVM optimization and a radial basis function (RBF) kernel. In a related study, (Kamarzarrin, Refan, Amiri, & Dameshghi, 2022) introduced a novel method for fault prognosis related to DFIG rotor winding, leveraging feature level fusion and adaptive thresholding based on process parameters. Their approach used classical time and frequency domain features to represent degradation behavior, with fault prognosis conducted using PSO-LS-SVM. Experimental validation, including simulated breakdown scenarios and comparisons with SVM- and NN-based approaches, showcases superior performance.

In their collective effort, the research group lead by Rezamand and Kordestani presented a comprehensive approach to predicting the RUL of WT generator bearings, addressing the challenges posed by varying operating conditions and uncertainty in prediction horizons. (Rezamand et al., 2021b) introduced a prognostic method integrating real-time SCADA data and vibration signals to assess the influence of environmental conditions on bearing failure dynamics, coupled with an adaptive Bayesian algorithm for RUL forecasting . In parallel, (Kordestani et al., 2022) proposed a feature extraction approach from vibration signals, followed by Bayesian RUL determination and high-level fusion methods such as the Hurwicz operator and Choquet integral to integrate RUL values and mitigate uncertainty.

2.3.3. Hybrid models

Only one work has been found aiming to predict RUL for WT generators with hybrid models, carried out by (Mehlan, Keller, & Nejad, 2023). A DT framework is introduced for the virtual sensing of WT hub loads. The research focuses on the estimation of aerodynamic hub loads, monitoring accumulated fatigue damage, and predicting the RUL of high-speed shaft generator side bearings. Using various data-driven regression models (linear regression (LR), SVR and tree ensembles) and a low-fidelity physics-based model, the bearing fatigue damage and RUL is based on ISO 281, which defines the equivalent dynamic load for cylindrical roller bearings. Then, long-term damage is obtained with Palmgren-Miner.

2.4. Other bearings

Eight works identified address the prediction of RUL of WT bearings that have not been previously listed, others have not been specified. All of them are included in this section. For instance, (Moynihan, Liberatore, Moaveni, & Khan, 2021) presented a physics-based approach to estimate RUL of main shaft bearings, employing strain measurements collected from blades and validated using real WT data, and fatigue life analysis is conducted using Miner's rule.

Within data-driven methods, (Jellali, Maatallah, & Ouni, 2022) proposed a method using temperature, viscosity, dynamic load, and fatigue damage for RUL prediction of WT main bearing, achieving high accuracy rates with a LR. The work of (Teng et al., 2020) emphasized a model-based approach using an improved unscented PF to study bearings located in the gearbox high-speed shaft, generator driven end, and generator non-driven end. This was done through measurement-centric methods, enhancing applicability for on-site WT scenarios. Moreover, (X. Li et al., 2023) built upon this foundation by introducing a method integrating degraded feature fusion models, threshold determination techniques, and self-constraint state-space estimator (SCSSE) to further enhance RUL prediction accuracy. (Encalada-Dávila, Puruncajas, Tutivén, & Vidal, 2021) developed an advanced prognostic approach, also ANN-based NBM, that relies solely on SCADA data to predict main bearing failure, enabling strategic maintenance scheduling several months in advance. Lastly, (Le, Lee, Dinh, & Park, 2024) compared similarity based model, employing an LSTM model, and degradation model, using LR and a stochastic exponential random model. The degradation models showed better performance. Lastly, (Bousebsi, Medoued, & Saidi, 2023) addressed RUL prediction for HSSB using the KS method. By incorporating Paris' Law and the KS, their method achieved enhanced accuracy in tracking degradation trends across five states.

Finally, (Yucesan & Viana, 2022) introduced the physics informed neural network (PiNN), a hybrid model for bearing fatigue damage accumulation. This was embedded as a RNN cell, where reduced-order physics models used for bearing fatigue damage accumulation, standardized bearing life formula found in ISO 281, and NNs represented grease degradation mechanism that quantifies grease damage that ultimately accelerates bearing fatigue.

2.5. Wind turbine as a system

Other six works have considered the drivetrain of a WT as a system. (Benmoussa, Djeziri, & Sanchez, 2020) proposed an integrated fault diagnosis and prognosis approach for WTs, employing a physics-based model, multi-class SVM classification, and a similarity-based method for RUL estimation without prior knowledge of degradation profiles. This method demonstrates effectiveness in handling uncertainties and transient operating modes, validated through a laboratory case study on a 750 kW WT. Similarly, (Binsbergen, Soares, Pedersen, & Nejad, 2022) developed a comprehensive physics- and SCADA-based model for RUL estimation in the WT drivetrain. By employing techniques such as load duration distribution and Miner's rule, their approach offers a holistic evaluation of the drivetrain's health and expected RUL.(de Souza Pereira Gomes et al., 2024) employed a RF model prediction, which sequentially conducts binary classification stages to determine if input samples represent conditions leading to

a failure event within a specified time period.

In their series of works, (L. Wang, Cao, Xu, & Liu, 2022) proposed innovative methodologies aimed at improving RUL estimation within the drivetrain of WTs. First, they introduced the gated graph convolutional network (GGCN), focusing on multi-sensor signal fusion and precise RUL prediction. By leveraging spatial-temporal graphs constructed from multi-sensor signals, the GGCN effectively captured both temporal and spatial dependencies crucial for understanding degradation states. Furthermore, (L. Wang, Cao, Ye, & Xu, 2023) addressed the need for uncertainty quantification by incorporating a quantile regression layer, providing confidence interval estimated essential for maintenance planning. Building upon this foundation, their subsequent work introduced the Bayesian large-kernel attention network (BLKAN) to enhance RUL prediction and uncertainty quantification for bearings. The BLKAN balanced computational efficiency with long-range correlations and channel adaptability, employing Bayesian large-kernel convolutions and variational inference to infer probability distributions of model parameters. Finally, (L. Wang, Cao, Ye, Xu, & Yan, 2024) presented the dual-view graph Transformer (DVGTformer), which enhances RUL prediction accuracy by fusing information from multiple sensors to capture complex degradation patterns. By integrating temporal and spatial perspectives through cascading layers of a graph transformer, the DVGTformer achieved superior performance compared to existing state-of-the-art methods.

3. Conclusions and future directions

This review illustrates the dynamic field of RUL estimation for WT components, showcasing the evolution and diversity of methodologies and their respective challenges. The comprehensive analysis of eighty-one papers published since 2018 on RUL prediction models reveals a clear alignment with the identified critical subsystems and failure modes within WT systems in Section 1. The predominant focus on predicting the RUL of gearboxes (35.8%), generators (32.1%), blades (14.8%), and associated bearings aligns the findings from the downtime analysis and failure costs in the introduction, where these components emerged as the most critical.

Physics-based, data-driven, and hybrid models have been identified to achieve an effective prognosis, all gathered in Figure 3. When the underlying physics of the system is known, e.g., bearing fatigue analysis with Miner´s rule, physics-based methods offer interpretable results and accuracy without the need for large amounts of data. It is difficult, though, to obtain robust physics-based models of these complex systems. Data-driven methods, while requiring less physical knowledge, can effectively quantify prognosis uncertainty and process high dimensional data, though they may lack interpretability and generalize poorly with limited data, a common occurrence in many WT datasets, where data availability is often sparse or low quality. Hybrid methods combine the advantages of various approaches, but may face challenges in model selection and characterization of uncertainty. Figure 2a illustrates that data-driven methodologies are predominant in predicting the RUL of WT components, with a growing inclination towards hybrid models.

While the primary objective of these models is to improve the accuracy and robustness of their predecessors, significant barriers remain. The challenges outlined in our review of RUL estimation methods for wind turbine components were derived from an extensive analysis of existing literature and research findings. Through a systematic review process, recurring obstacles were identified and categorized into five distinct groups: inherent uncertainty management, integration of physical knowledge, consideration of variable operational conditions, data issues and complex system dynamics. This classification was based on the underlying nature of the challenges and their impact on prognostic accuracy and reliability. The distribution of papers that address these issues is shown in Figure 4 (one paper can address more than one challenge). These are discussed below.

1. **Uncertainty** from various sources, such as sensor noise and model variability, is a key obstacle. Thus, its quantification is essential to reduce the impact of uncertainties throughout maintenance optimization and decision-making. Techniques such as Bayesian belief networks (BBNs) and Monte Carlo methods aimed at minimizing it.
2. **Integration of physical knowledge**. The integration of underlying physics of the system is gaining attention to obtain higher accuracy and credibility of the prognosis, as shown in the hybrid methods subsections. However, there remains a need for further research and enhancement. Methods to integrate physical knowledge into NNs such as PiNNs offer promising avenues for accurate and interpretable prognostics (Chen, Ma, Zhao, Zhai, & Mao, 2022). In the industrial application facet, RUL prediction techniques have been applied to a number of important fields, including but not limited to aerospace systems, industrial robots, wind power systems, high-speed trains, etc. (H. Li, Zhang, Li, & Si, 2024), but there still remains a significant challenge in fully implementing these techniques within the components of WTs. However, there are some key limitations and challenges in model property aspects, which (Xu, Kohtz, Boakye, Gardoni, & Wang, 2023) summarized into five types: model selection, model structure, model parameter, model optimizer, and model prediction.
3. **Variable operational conditions**. WTs operate in dynamic environments characterized by fluctuating wind speeds, changing loads, and varying environmental con-

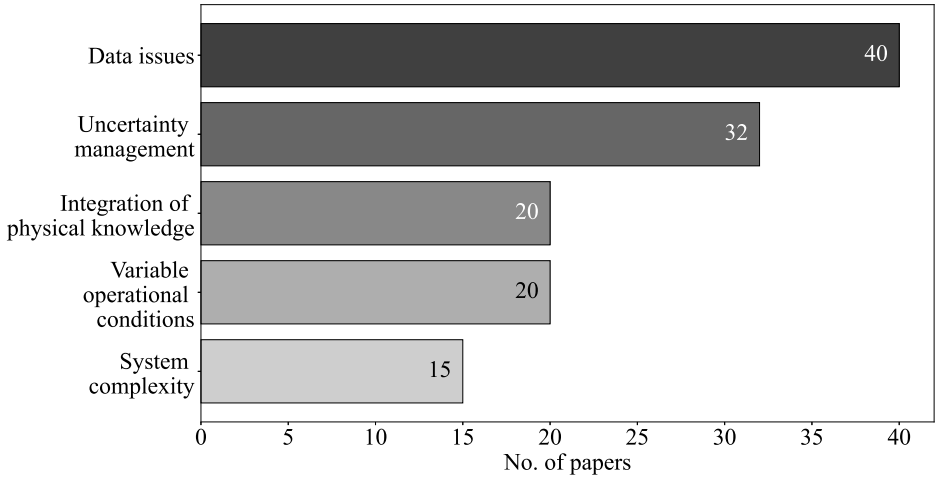


Figure 4. Amount of papers addressing challenges in WT component RUL prediction.

ditions. Such variability not only impacts the performance and wear of individual components, but also complicates the prediction of their future reliability and the accuracy of RUL models, often trained with similar operating conditions. The need for continuous online learning underscores the urgency for novel algorithms and hardware architectures (Elattar, El-Brawany, Elminir, Ibrahim, & Ramadan, 2023). In this context, transfer learning and domain adaptation present an opportunity to adapt models to varying operating conditions, although challenges remain to ensure prediction accuracy across different equipment (Ramezani et al., 2023). Therefore, the transferability assessment of different domains continues to pose a substantial challenge.

4. **Data issues**. Noisy measurements and a notorious lack of high-quality data are among these challenges. These difficulties are compounded by the scarcity of samples and monitoring data, which makes accurate predictive modeling even more challenging. Moreover, in many cases, only SCADA data are available, which presents a notable barrier, aggravated by the limited and low frequency samples, and the unbalanced nature of the condition data available for analysis. To overcome these obstacles, researchers must navigate the complexities of fusing multi-sensor signals to enrich the available data sources. Several papers within this study employ methodologies for multi-signal feature extraction and dimensionality reduction, such as PCA. However, it is beyond the scope of this review to delve into these techniques.
5. **System complexity**. WT are inherently complex systems, characterized by a multitude of interconnected components, as mentioned in the introduction. On one hand, multiple faults in a single component are a frequent occurrence, which are not considered in academic studies. On the other hand, the degradation process should take into account the way different components interact. To gain interpretability, more signal processing procedures could also be applied to the machine degradation process.

As research progresses, it is essential to address the identified challenges systematically, paying particular attention to the critical components within WT components. The implementation of PHM will facilitate the development of a robust predictive maintenance plan, which will contribute to the reduction of LCOE associated with O&M costs, thus aligning with the objectives of achieving green transition goals.

ACKNOWLEDGMENT

This study was partially funded by the Spanish Agencia Estatal de Investigación (AEI) — Ministerio de Ciencia e Innovación, the Fondo Europeo de Desarrollo Regional (FEDER) through the research projects PID2021-122132OB-C21, PID 2021-122132OB-C22, and TED2021-129512B-I00; and by the Generalitat de Catalunya through the research project 2021-SGR-01044.

REFERENCES

Benmoussa, S., Djeziri, M. A., & Sanchez, R. (2020). Support vector machine classification of current data for fault diagnosis and similarity-based approach for failure prognosis in wind turbine systems. In M. Sayed-Mouchaweh (Ed.), *Artificial intelligence techniques for a scalable energy transition: Advanced methods, digital technologies, decision support tools, and applications* (pp. 157–182). Cham: Springer International Publishing.

Bermúdez, K., Ortiz-Holguin, E., Tutivén, C., Vidal, Y., & Benalcázar-Parra, C. (2022). Wind turbine main bearing failure prediction using a hybrid neural net-

work. *Journal of Physics: Conference Series*, *2265*(3), 032090.

Binsbergen, D. V., Soares, M. N., Pedersen, E., & Nejad, A. R. (2022). A physics-, scada-based remaining useful life calculation approach for wind turbine drivetrains. *Journal of Physics: Conference Series*, *2265*(3), 032079.

Bousebsi, M., Medoued, A., & Saidi, L. (2023). Data-driven and physics-based approaches for wind turbine high-speed shaft bearing prognostics. In *Proceedings of 2023 international conference on control, automation and diagnosis (iccad)* (p. 1-6). IEEE.

Boutrous, K., Puig, V., & Nejjari, F. (2022). A set-based prognostics approach for wind turbine blade health monitoring. *IFAC-PapersOnLine*, *55*(6), 402-407.

Cao, L., Qian, Z., & Pei, Y. (2019). Remaining useful life prediction of wind turbine generator bearing based on emd with an indicator. In *Proceedings of prognostics and system health management conference, phm-chongqing* (p. 375-379). IEEE.

Cao, L., Qian, Z., Zareipour, H., Wood, D., Mollasalehi, E., Tian, S., & Pei, Y. (2018). Prediction of remaining useful life of wind turbine bearings under non-stationary operating conditions. *Energies*, *11*(12), 3318.

Cao, L., Zhang, H., Meng, Z., & Wang, X. (2023). A parallel gru with dual-stage attention mechanism model integrating uncertainty quantification for probabilistic rul prediction of wind turbine bearings. *Reliability Engineering and System Safety*, *235*, 109197.

Carroll, J., Koukoura, S., McDonald, A., Charalambous, A., Weiss, S., & McArthur, S. (2019). Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques. *Wind Energy*, *22*(3), 360-375.

Catelani, M., Ciani, L., Galar, D., & Patrizi, G. (2020, 9). Optimizing maintenance policies for a yaw system using reliability-centered maintenance and data-driven condition monitoring. *IEEE Transactions on Instrumentation and Measurement*, *69*(9), 6241-6249.

Chen, X., Ma, M., Zhao, Z., Zhai, Z., & Mao, Z. (2022). Physics-informed deep neural network for bearing prognosis with multisensory signals. *Journal of Dynamics, Monitoring and Diagnostics*, *1*(4), 200-207.

Cheng, F., Qu, L., & Qiao, W. (2018). Fault prognosis and remaining useful life prediction of wind turbine gearboxes using current signal analysis. *IEEE Transactions on Sustainable Energy*, *9*(1), 157-167.

Cheng, F., Qu, L., Qiao, W., & Hao, L. (2019). Enhanced particle filtering for bearing remaining useful life prediction of wind turbine drivetrain gearboxes. *IEEE Transactions on Industrial Electronics*, *66*(6), 4738-4748.

Dameshghi, A., & Refan, M. H. (2021). Combination of condition monitoring and prognosis systems based on current measurement and pso-ls-svm method for wind turbine dfigs with rotor electrical asymmetry. *Energy Systems*, *12*(1), 203-232.

Dao, C., Kazemtabrizi, B., & Crabtree, C. (2019, 12). Wind turbine reliability data review and impacts on levelised cost of energy. *Wind Energy*, *22*, 1848-71.

Desai, A., Guo, Y., Sheng, S., Phillips, C., & Williams, L. (2020). Prognosis of wind turbine gearbox bearing failures using scada and modeled data. In *Annual conference of the phm society* (Vol. 12, p. 10).

de Souza Pereira Gomes, G., de Andrade Lopes, S. M., Araujo, D. C. P., Flauzino, R. A., Pinto, M. M., & Alves, M. E. G. (2024). Wind turbine remaining useful life prediction using small dataset and machine learning techniques. *Journal of Control, Automation and Electrical Systems*, *35*, 337-345.

Du, X., Jia, W., Yu, P., Shi, Y., & Gong, B. (2023). Rul prediction based on gam–cnn for rotating machinery. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *45*(3), 142.

Elasha, F., Shanbr, S., Li, X., & Mba, D. (2019). Prognosis of a wind turbine gearbox bearing using supervised machine learning. *Sensors*, *19*(14), 3092.

Elattar, H. M., El-Brawany, M. A., Elminir, H. K., Ibrahim, D. A., & Ramadan, E. (2023, Oct.). Artificial intelligence-based data-driven prognostics in industry: A survey. *Computers & Industrial Engineering*, *184*, 109605.

Elforjani, M. (2020). Diagnosis and prognosis of real world wind turbine gears. *Renewable Energy*, *147*(1), 1676-1693.

Encalada-Dávila, A., Puruncajas, B., Tutivén, C., & Vidal, Y. (2021). Wind turbine main bearing fault prognosis based solely on scada data. *Sensors*, *21*(6), 2228.

Farhat, M. H., Chaari, F., Chiementin, X., Bolaers, F., & Haddar, M. (2022). Dynamic remaining useful life estimation for a shaft bearings system. In F. Chaari, X. Chiementin, R. Zimroz, F. Bolaers, & M. Haddar (Eds.), *Smart monitoring of rotating machinery for industry 4.0* (p. 169-178). Cham: Springer International Publishing.

Ferreira, C., & Gonçalves, G. (2022, Apr.). Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems*, *63*, 550-562.

Guo, Y., Sheng, S., Phillips, C., Keller, J., Veers, P., & Williams, L. (2020, July). A methodology for reliability assessment and prognosis of bearing axial cracking in wind turbine gearboxes. *Renewable and Sustainable Energy Reviews*, *127*, 109888.

Hayder, A. N., & Saidi, L. (2021). Applications of artificial neural networks with input and output degradation data for renewable energy systems fault prognosis. In *2021 12th international renewable energy congress (irec)* (p. 1-6).

He, K., Su, Z., Tian, X., Yu, H., & Luo, M. (2022, Jan.). Rul prediction of wind turbine gearbox bearings based on self-calibration temporal convolutional network. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-12.

Hu, Y., Li, H., Shi, P., Chai, Z., Wang, K., Xie, X., & Chen, Z. (2018, Nov.). A prediction method for the real-time remaining useful life of wind turbine bearings based on the wiener process. *Renewable Energy*, *127*, 452-460.

IREA. (2023). *Renewable power generation costs 2022.*

ISO. (2012). *Condition monitoring and diagnostics of machines-vocabulary* (Tech. Rep. No. 13372:2012). International Organization for Standardization.

Jaramillo, F., Gutiérrez, J. M., Orchard, M., Guarini, M., & Astroza, R. (2022, July). A bayesian approach for fatigue damage diagnosis and prognosis of wind turbine blades. *Mechanical Systems and Signal Processing*, *174*, 109067.

Jellali, A., Maatallah, H., & Ouni, K. (2022). Predicting remaining useful life of wind turbine bearing using linear regression. In *2022 5th international conference on advanced systems and emergent technologies (ic_aset)* (p. 357-362).

Jiang, Z., Hu, W., Dong, W., Gao, Z., & Ren, Z. (2017). Structural reliability analysis of wind turbines: A review. *Energies*, *10*(12), 2099.

Kamarzarrin, M., Refan, M. H., Amiri, P., & Dameshghi, A. (2022). A new intelligent fault diagnosis and prognosis method for wind turbine doubly-fed induction generator. *Wind Engineering*, *46*(1), 308-340.

Kordestani, M., Rezamand, M., Orchard, M. E., Carriveau, R., Ting, D. S., Rueda, L., & Saif, M. (2022). New condition-based monitoring and fusion approaches with a bounded uncertainty for bearing lifetime prediction. *IEEE Sensors Journal*, *22*(9), 9078-9086.

Kramti, S. E., Ali, J. B., Saidi, L., Sayadi, M., Bouchouicha, M., & Bechhoefer, E. (2021). A neural network approach for improved bearing prognostics of wind turbine generators. *EPJ Applied Physics*, *93*(2), 20901.

Kramti, S. E., Ben Ali, J., Saidi, L., Sayadi, M., & Bechhoefer, E. (2018). Direct wind turbine drivetrain prognosis approach using elman neural network. In *2018 5th international conference on control, decision and information technologies (codit)* (p. 859-864).

Kramti, S. E., Saidi, L., Ali, J. B., Sayadi, M., & Bechhoeferer, E. (2019). Particle filter based approach for wind turbine high-speed shaft bearing health prognosis. In *2019 international conference on signal, control and communication (scc)* (p. 46-50).

Lan, X., Liang, D., Liu, W., Wang, Y., Han, Y., & Yu, Z. (2023). Remaining useful life prediction of wind turbine generator bearings based on nonlinear wiener process. In *Proceedings - 2023 3rd power system and green energy conference, psgec 2023* (p. 1124-1129). Institute of Electrical and Electronics Engineers Inc.

Le, T.-T., Lee, S.-J., Dinh, M.-C., & Park, M. (2024). Design of an improved remaining useful life prediction model based on vibration signals of wind turbine rotating components. *Energies*, *17*(1), 19.

Lee, I. Y., Roh, H. D., & Park, Y. B. (2022, Nov.). Prognostics and health management of composite structures under multiple impacts through electromechanical behavior and a particle filter. *Materials and Design*, *223*, 111143.

Li, B., Tang, B., Deng, L., & Zhao, M. (2021, June). Self-attention convlstm and its application in rul prediction of rolling bearings. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-11.

Li, H., Zhang, Z., Li, T., & Si, X. (2024). A review on physics-informed data-driven remaining useful life prediction: Challenges and opportunities. *Mechanical Systems and Signal Processing*, *209*(111120). doi: 10.1016/j.ymssp.2024.111120

Li, X., Teng, W., Peng, D., Ma, T., Wu, X., & Liu, Y. (2023, May). Feature fusion model based health indicator construction and self-constraint state-space estimator for remaining useful life prediction of bearings in wind turbines. *Reliability Engineering and System Safety*, *233*, 109124.

Li, Z., Zhang, X., Kari, T., & Hu, W. (2021). Health assessment and remaining useful life prediction of wind turbine high-speed shaft bearings. *Energies*, *14*(15), 4612.

Liu, M., Dong, Z., & Shi, H. (2023). Multi-sensor information fusion and multi-model fusion-based remaining useful life prediction of fan slewing bearings with the nonlinear wiener process. *Sustainability (Switzerland)*, *15*(15), 12010.

Liu, Z., & Zhang, L. (2020, Jan.). A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement: Journal of the International Measurement Confederation*, *149*, 107002.

Lydia, M., & Edwin Prem Kumar, G. (2023). 10 - condition monitoring in wind turbines: a review. In F. P. Garcia Marquez, M. Papaelias, & V. L. J. Junior (Eds.), *Non-destructive testing and condition monitoring techniques in wind energy* (p. 229-247). Academic Press.

Lázaro, R., Yürüen, N. Y., & Melero, J. J. (2020). Determining remaining lifetime of wind turbine gearbox using a health status indicator signal. *Journal of Physics: Conference Series*, *1618*(2), 022037.

Mehlan, F. C., Keller, J., & Nejad, A. R. (2023). Virtual sensing of wind turbine hub loads and drivetrain fatigue damage. *Forschung im Ingenieurwesen/Engineering Research*, *87*(1), 207-218.

Mehlan, F. C., Nejad, A. R., & Gao, Z. (2022). Digital twin based virtual sensor for online fatigue damage mon-

itoring in offshore wind turbine drivetrains. *Journal of Offshore Mechanics and Arctic Engineering*, *144*(6), 060901.

Merainani, B., Laddada, S., Bechhoefer, E., Chikh, M. A. A., & Benazzouz, D. (2022). An integrated methodology for estimating the remaining useful life of high-speed wind turbine shaft bearings with limited samples. *Renewable Energy*, *182*, 1141-1151.

Moroney, P. D., & Verma, A. S. (2023). Durability and damage tolerance analysis approaches for wind turbine blade trailing edge life prediction: A technical review. *Energies*, *16*(24), 7934.

Moynihan, B., Liberatore, S., Moaveni, B., & Khan, U. (2021). Fatigue life analysis of main shaft bearings in wind turbines using strain measurements collected on blades. In (p. 185-192). Springer.

Muto, K., Namura, N., Ukei, Y., & Takeda, N. (2019). Model-based load estimation for wind turbine blade with kalman filter. In *2019 8th international conference on renewable energy research and applications (icrera)* (p. 191-199).

Olabi, A. G., Wilberforce, T., Elsaid, K., Sayed, E. T., Salameh, T., Abdelkareem, M. A., & Baroutaji, A. (2021, 9). *A review on failure modes of wind turbine components* (Vol. 14) (No. 17). MDPI AG.

Owolabi, O. I., Madushele, N., Adedeji, P. A., & Olatunji, O. O. (2023, 12). *Fem and ann approaches to wind turbine gearbox monitoring and diagnosis: a mini review* (Vol. 9) (No. 4). Springer Science and Business Media Deutschland GmbH.

Pagitsch, M., Jacobs, G., & Bosse, D. (2020). Remaining useful life determination for wind turbines. *Journal of Physics: Conference Series*, *1452*(1), 012052.

Pan, Y., Hong, R., Chen, J., & Wu, W. (2020, June). A hybrid dbn-som-pf-based prognostic approach of remaining useful life for wind turbine gearbox. *Renewable Energy*, *152*, 138-154.

Pandit, R., & Xie, W. (2023). Data-driven models for predicting remaining useful life of high-speed shaft bearings in wind turbines using vibration signal analysis and sparrow search algorithm. *Energy Science and Engineering*, *11*(1), 4557-4569.

Peng, C., Chen, Q., Zhou, X., Wang, S., & Tang, Z. (2020). Wind turbine blades icing failure prognosis based on balanced data and improved entropy. *Int. J. Sen. Netw.*, *34*(2), 126–135.

Peng, Y., Bi, R., & Wang, Y. (2023). Regime-switching model with adaptive adjustments for degradation prognosis. *IEEE Transactions on Instrumentation and Measurement*, *72*(6), 1-11.

Peter, R., Zappalá, D., Schamboeck, V., & Watson, S. J. (2022). Wind turbine generator prognostics using field scada data. *Journal of Physics: Conference Series*, *2265*(3), 032111.

Praveen, H. M., Shah, D., Pandey, K. D., Vamsi, I., & Sabareesh, G. R. (2019). Pca based health indicator for remaining useful life prediction of wind turbine gearbox. In *Vibroengineering procedia* (Vol. 29, p. 31-36). EXTRICA.

Qiao, W., & Qu, L. (2018). Prognostic condition monitoring for wind turbine drivetrains via generator current analysis. *Chinese Journal of Electrical Engineering*, *4*(3), 80-89.

Qin, Y., Yang, J., Zhou, J., Pu, H., & Mao, Y. (2023, Apr.). A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery rul prediction. *Advanced Engineering Informatics*, *56*, 101973.

Ramezani, S. B., Cummins, L., Killen, B., Carley, R., Amirlatifi, A., Rahimi, S., ... Bian, L. (2023). Scalability, explainability and performance of data-driven algorithms in predicting the remaining useful life: A comprehensive review. *IEEE Access*, *11*(4), 41741-41769.

Rezamand, M., Carriveau, R., Ting, D. S., Davison, M., & Davis, J. J. (2019). Aggregate reliability analysis of wind turbine generators. *IET Renewable Power Generation*, *13*(11), 1902-1910.

Rezamand, M., Kordestani, M., Carriveau, R., Ting, D. S., & Saif, M. (2020, Mar.). An integrated feature-based failure prognosis method for wind turbine bearings. *IEEE/ASME Transactions on Mechatronics*, *25*, 1468-1478.

Rezamand, M., Kordestani, M., Orchard, M., Carriveau, R., Ting, D., & Saif, M. (2021a). Condition monitoring and failure prognostic of wind turbine blades. In *2021 ieee international conference on systems, man, and cybernetics (smc)* (p. 1711-1718).

Rezamand, M., Kordestani, M., Orchard, M. E., Carriveau, R., Ting, D. S., & Saif, M. (2021b, Mar.). Improved remaining useful life estimation of wind turbine drivetrain bearings under varying operating conditions. *IEEE Transactions on Industrial Informatics*, *17*, 1742-1752.

Saathoff, M., Rosemeier, M., Kleinselbeck, T., & Rathmann, B. (2021). Effect of individual blade pitch angle misalignment on the remaining useful life of wind turbines. *Wind Energy Science*, *6*(5), 1079-1087.

Saidi, L., Ali, J. B., Benbouzid, M., & Bechhofer, E. (2018, Sep.). An integrated wind turbine failures prognostic approach implementing kalman smoother with confidence bounds. *Applied Acoustics*, *138*, 199-208.

Salameh, J. P., Cauet, S., Etien, E., Sakout, A., & Rambault, L. (2018, Oct.). Gearbox condition monitoring in wind turbines: A review. *Mechanical Systems and Signal Processing*, *111*, 251-264.

Shen, Y., Tang, B., Li, B., Tan, Q., & Wu, Y. (2022, Oct.). Remaining useful life prediction of rolling bearing based on multi-head attention embedded bi-lstm

network. *Measurement: Journal of the International Measurement Confederation*, *202*, 111803.

Song, M., Youliang, S., Kai, J., Cheng, L., & Tao, W. (2020). Remaining life prediction of wind turbine bearing based on wiener process. *IOP Conference Series: Materials Science and Engineering*, *788*(1), 012089.

Tayade, A., Patil, S., Phalle, V., Kazi, F., & Powar, S. (2019). Remaining useful life (rul) prediction of bearing by using regression model and principal component analysis (pca) technique. In (Vol. 23, p. 30-36). EXTRICA.

Tazi, N., Châtelet, E., & Bouzidi, Y. (2017). Using a hybrid cost-fmea analysis for wind turbine reliability analysis. *Energies*, *10*, 276.

Teng, W., Han, C., Hu, Y., Cheng, X., Song, L., & Liu, Y. (2020, Mar.). A robust model-based approach for bearing remaining useful life prognosis in wind turbines. *IEEE Access*, *8*, 47133-47143.

Valeti, B., & Pakzad, S. N. (2018). Remaining useful life estimation of wind turbine blades under variable wind speed conditions using particle filters. In *Annual conference of the phm society* (Vol. 10).

Valeti, B., & Pakzad, S. N. (2019). Estimation of remaining useful life of a fatigue damaged wind turbine blade with particle filters. In S. Pakzad (Ed.), *Dynamics of civil structures, volume 2* (p. 319-328). Cham: Springer International Publishing.

Verma, A., Zappalá, D., Sheng, S., & Watson, S. J. (2022). Wind turbine gearbox fault prognosis using high-frequency scada data. *Journal of Physics: Conference Series*, *2265*(3), 032067.

Vettori, S., Lorenzo, E. D., Peeters, B., Luczak, M. M., & Chatzi, E. (2023, Feb.). An adaptive-noise augmented kalman filter approach for input-state estimation in structural dynamics. *Mechanical Systems and Signal Processing*, *184*, 109654.

Wang, J., Gao, R. X., Yuan, Z., Fan, Z., & Zhang, L. (2019). A joint particle filter and expectation maximization approach to machine condition prognosis. *Journal of Intelligent Manufacturing*, *30*(2), 605-621.

Wang, J., Liang, Y., Zheng, Y., Gao, R. X., & Zhang, F. (2020, Jan.). An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renewable Energy*, *145*, 642-650.

Wang, L., Cao, H., Xu, H., & Liu, H. (2022). A gated graph convolutional network with multi-sensor signals for remaining useful life prediction. *Knowledge-Based Systems*, *252*, 109340.

Wang, L., Cao, H., Ye, Z., & Xu, H. (2023, Oct.). Bayesian large-kernel attention network for bearing remaining useful life prediction and uncertainty quantification. *Reliability Engineering and System Safety*, *238*, 109421.

Wang, L., Cao, H., Ye, Z., Xu, H., & Yan, J. (2024, Jan.). Dvgtformer: A dual-view graph transformer to fuse multi-sensor signals for remaining useful life prediction. *Mechanical Systems and Signal Processing*, *207*, 110935.

Wang, P., Long, Z., & Wang, G. (2020). A hybrid prognostics approach for estimating remaining useful life of wind turbine bearings. *Energy Reports*, *6*(9), 173-182.

Wang, Z., Gao, P., & Chu, X. (2022). Remaining useful life prediction of wind turbine gearbox bearings with limited samples based on prior knowledge and pi-lstm. *Sustainability (Switzerland)*, *14*(19), 12094.

Xiang, S., Li, P., Luo, J., & Qin, Y. (2024). Micro transfer learning mechanism for cross-domain equipment rul prediction. *IEEE Transactions on Automation Science and Engineering*, 1-11. doi: 10.1109/TASE.2024.3366288

Xiang, S., Qin, Y., Liu, F., & Gryllias, K. (2022, July). Automatic multi-differential deep learning and its application to machine remaining useful life prediction. *Reliability Engineering and System Safety*, *223*, 108531.

Xiang, S., Qin, Y., Luo, J., & Pu, H. (2022). Spatiotemporally multidifferential processing deep neural network and its application to equipment remaining useful life prediction. *IEEE Transactions on Industrial Informatics*, *18*(10), 7230-7239.

Xiang, S., Qin, Y., Luo, J., Wu, F., & Gryllias, K. (2023). A concise self-adapting deep learning network for machine remaining useful life prediction. *Mechanical Systems and Signal Processing*, *191*, 110187.

Xu, Y., Kohtz, S., Boakye, J., Gardoni, P., & Wang, P. (2023). Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges. *Reliability Engineering and System Safety*, *230*(108900). doi: 10.1016/j.ress.2022.108900

Yucesan, Y. A., & Viana, F. A. (2022, May). A hybrid physics-informed neural network for main bearing fatigue prognosis under grease quality variation. *Mechanical Systems and Signal Processing*, *171*, 108875.

Yue, G., Ping, G., & Lanxin, L. (2018). An end-to-end model based on cnn-lstm for industrial fault diagnosis and prognosis. In *2018 international conference on network infrastructure and digital content (ic-nidc)* (p. 274-278).

Zheng, H., Deng, W., Song, W., Cheng, W., Cattani, P., & Villecco, F. (2024). Remaining useful life prediction of a planetary gearbox based on meta representation learning and adaptive fractional generalized pareto motion. *Fractal and Fractional*, *8*(1), 14.

BIOGRAPHIES

Jokin Cuesta holds a degree in physics by the University of the Basque Country - Euskal Herriko Unibertsitatea (EHU-UPV, 2020). He furthered his education with a

focus on data science master at Universitat Oberta de Catalunya (UOC, 2023) in Barcelona. He is currently engaged in the pursuit of an industrial PhD at the Artificial Intelligence and Data area in Ikerlan and the Control, Data, and Artificial Intelligence (CoDAlab) group in the Department of Mathematics at Universitat Politècnica de Catalunya. With a passion for the intersection of AI and CM, his research is dedicated to developing innovative methodologies for CM in the wind energy sector. His specific focus lies in identifying degradation patterns in WT components and estimating their RUL.

Urko Leturiondo is Industrial Engineer majored in Mechanics and Industrial Design by Tecnun - Engineering School of Universidad de Navarra (2012) and he received his PhD degree by Luleå University of Technology (2016). In 2012 he joined the Mechanical Engineering Department of Ikerlan as a PhD student, doing his research in collaboration with the Division of Operation and Maintenance Engineering of Luleå University of Technology. From 2016 to 2019 he worked as a researcher in the Control and Monitoring Area of Ikerlan. Since 2019 he has led different research teams in Ikerlan, being currently the DataOps team leader. His current research interests include condition monitoring, asset management, modelling of mechatronic systems, vibration and acoustic analysis, data analytics, artificial intelligence and DevOps.

Yolanda Vidal holds a degree in Mathematics (1999) and a PhD in Applied Mathematics (2005) from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. As an Associate Professor at UPC and an IEEE Senior Member, she actively engages in multidisciplinary research. Her areas of expertise include condition monitoring, structural health monitoring, fault diagnosis and prognosis, predictive maintenance, machine/deep learning, mathematical modeling, and the application of these disciplines in WT technologies. She serves on the Editorial Board of several international journals, including Engineering Applications of AI (Elsevier), Wind Energy (Wiley), Wind Energy Science (Copernicus), Mathematics, Sensors, Energies, Frontiers in Built Environment, and Frontiers in Energy Research. Her prolific contributions are evidenced by more than 60 high-impact journal articles, 20 competitive R+D+I projects, 17 book chapters, 10 books, 3 supervised PhD theses, 1 invention patent, a collaboration contract with an industry partner, and over 110 conference papers.

Francesc Pozo obtained his degree in mathematics from the University of Barcelona in 2000, and his PhD in applied mathematics from the Universitat Politècnica de Catalunya (UPC) in 2005. Since 2000, he has been with the Department of Mathematics at UPC, where he is now a Full Professor and the coordinator of the Control, Data, and Artificial Intelligence research group. His expertise encompasses condition monitoring, control systems, data-driven modeling, identification, and structural health monitoring, with a special focus on WTs. He is an Editorial Board Member for several international journals, such as Structural Control and Health Monitoring, International Journal of Distributed Sensor Networks, Mathematical Problems in Engineering, Mathematics, Sensors, Algorithms, Journal of Vibration and Control, Frontiers in Built Environment, Frontiers in Energy Research, and Energies. His contributions to his field include over 70 high-impact journal articles, participation in 23 competitive R&D&I projects, authorship of 34 book chapters and 12 books, mentorship of 6 PhD candidates, the filing of 1 invention patent, a collaboration contract with an industry partner, and more than 130 conference papers.

A rolling bearing state evaluation method based on deep learning combined with Wiener process

Yuntian Ta[1], Tiantian Wang[2], Jingsong Xie[3], Jinsong Yang[4], and Tongyang Pan[5]

[1,2,3,4,5]*School of Traffic & Transportation Engineering, Central South University, Changsha, China*

234201008@csu.edu.cn
wangtiantian@csu.edu.cn
jingsongxie@foxmail.com
yangjs@csu.edu.cn
typ2022@163.com

ABSTRACT

As a key component of rotating parts, rolling bearings largely determine the operation safety of equipment. However, in practical applications, because the degradation trajectory of rolling bearings cannot be truly characterized, the existing model cannot accurately describe the degradation trajectory of rolling bearings, resulting in the running state of rolling bearings cannot be directly evaluated. Therefore, a method of rolling bearing state assessment based on deep learning combined with Wiener process is proposed in this paper. Firstly, a deep network model is constructed by deep learning to mine the degradation information of rolling bearings. Secondly, the mined degradation information is fused, and then the degradation indicator used to characterize the degraded trajectory of the rolling bearing is constructed. Then, based on Wiener process, the degradation model of rolling bearing is established to describe the degradation mode of rolling bearing. Finally, the constructed degradation indicator is input into the established degradation model to predict its RUL, and then the running state of the rolling bearing is evaluated.

1. INTRODUCTION

During the operation of mechanical equipment, due to the influence of many factors, mechanical equipment will inevitably degrade. This degradation process generally occurs first in components that produce relative motion, especially rolling bearings(Zhu et al. 2024). Therefore, in order to ensure that mechanical equipment always serves in a safe state, it is very necessary to evaluate the operating status of rolling bearings. The remaining useful life (RUL) prediction method has been recognized as a basic and effective method for state assessment of rolling bearings(Li et al. 2024). (If the RUL of the rolling bearing can be predicted, the current service status of the rolling bearing can be assessed) Currently, in the field of prediction of the RUL of rolling bearings, scholars have proposed a series of life prediction methods of rolling bearings, but generally they can be divided into methods based on expert knowledge base, data-driven, physical models and hybrid methods(Wang et al. 2023).

The method based on expert knowledge base achieves prediction by comparing the similarity between the observed data and the previously defined fault database through expert system or fuzzy system(Qin et al. 2023). For example, Qin et al. proposed a two-stage RUL prediction method based on similarity, constructing a degradation indicator (DI) of bearings through a multi-head self-attention mechanism, and comparing the constructed DI with other bearing degradation indexes in the expert knowledge base, thereby realizing the prediction of the RUL of the bearing(Qin et al. 2023). Xia et al. proposed a hybrid Gaussian-evidence hidden Markov model that integrates expert knowledge and condition monitoring information to predict the RUL of bearings under the framework of belief function theory(Xiahou, Zeng, and Liu 2021). These methods often require special knowledge about the fault data, however obtaining this knowledge is expensive in practice. The data-driven method uses the historical status data of the equipment to extract characteristic information related to the status changes of the monitored object. Through statistical analysis, pattern recognition, machine learning and other technologies, it attempts to simulate the fuzzy functional relationship between sensor data and equipment status, and then realize the status assessment and RUL prediction of the monitored object(Li et al. 2022). For example, Cheng et al. extracted nonlinear features from bearing vibration signals and inputted them into convolutional neural networks to evaluate the health status of bearings, and combined them with relevant vector machines to predict the RUL of bearings(Cheng et al. 2021). Yoo et al. used continuous wavelet transform to convert bearing vibration signals into image signals and input them into convolutional neural networks for predicting the RUL of bearings(Yoo and Baek 2018). Ren et al. used deep self-coding neural networks to compress the time-frequency wavelet features of rolling bearings and predict the RUL of

rolling bearings(Ren et al. 2018). However, these methods need to establish the state characterization function of the rolling bearing, and with the increase of the prediction time span, the characterization ability of the model decreases, and the prediction accuracy of the RUL decreases. The physical model method is based on the mathematical representation of the physical behavior during the degradation process to predict the degradation performance and RUL of the bearing. For example, Kogan et al. established a multi-body dynamics model of rolling bearings based on classical dynamics and kinematic equations to describe the health degradation process of rolling bearings under different faults and predict their RUL by fitting its degradation process(Kogan et al. 2015). Qian et al. improved the Paris-Erdogan model and constructed a multi-time scale degradation model to track the changes in the degradation rate of the bearing in different time periods to predict the RUL of the bearing(Qian, Yan, and Gao 2017). These methods can provide accurate prediction results, it still requires an in-depth understanding of the physical characteristics of the bearing and the prognosis of the bearing. The accuracy depends heavily on the accuracy of the physical model used. The hybrid prediction method is a RUL prediction method that combines the advantages of physical models and data drivers(Wang et al. 2020). Wang et al. constructed a new scalable two-stage linear/nonlinear composite model to describe various degradation behaviors of bearings through a hybrid data- and model-driven method, and predicted the RUL of bearings by using a long and short time memory network(Wang, Cui, and Wang 2022). Rezamand et al. defined the role of environmental conditions in the dynamics of bearing failure. They achieved the RUL prediction of faulty bearings through vibration signal recognition and fault dynamics analysis(Rezamand et al. 2021). The hybrid prediction method can effectively simulate the degradation process of rolling bearings. However, these methods complicate the algorithm and is limited by the physical behavior of the rolling bearing during the degradation process, which in turn leads to modeling difficulties.

Due to the limitations of different methods, the unclear exploration of the failure mechanism of rolling bearings, the lack of degradation data, and especially the neglect of historical operating data of rolling bearings in normal service, these methods cannot accurately evaluate the service status of rolling bearings. There are two reasons for this. First, the degradation characteristics used cannot accurately represent the degradation trajectory of rolling bearings; second, the degradation model used cannot map the failure mechanism of rolling bearings. Due to the powerful feature extraction ability of convolutional neural networks, by stacking multiple convolutional and pooling layers, more and more abstract and advanced features can be gradually extracted. This hierarchical feature extraction can better capture the degradation information of bearings, thereby improving the performance of the model. In addition, due to the excellent non monotonic characteristics of the Wiener process, it can effectively describe the local fluctuation characteristics on the degradation path of bearings. Therefore, in order to overcome the limitations of the above methods, this paper proposes a rolling bearing state assessment method based on deep learning combined with Wiener process, starting from the construction of degradation indicators of rolling bearings and the failure mechanism mapping of the model. This method first constructs a degradation indicator extractor for the full- life cycle of rolling bearings based on one-dimensional convolutional neural. Secondly, a mapping model between its degradation trajectory and RUL is established based on the Wiener process. Then, using DI to estimate the unknown parameters in the model, the RUL prediction of the rolling bearing at different monitoring points is completed. Finally, the status evaluation of the rolling bearing is realized through the prediction results at the current moment.

2. METHOD PROPOSED

2.1. DI construction method

Convolutional neural network is a type of deep neural network, which consists of multiple neural network layers. Each layer consists of multiple neurons that are connected to the neurons in the previous layer. Convolutional neural networks usually contain three types of layers: convolutional layers, pooling layers, and fully connected layers. Because the dimensional convolutional neural network has good information mining and weight sharing capabilities(She and Jia 2019). Therefore, this paper constructs the bearing degradation index of the rolling shaft based on the one-dimensional convolutional neural network. The specific construction method is as follows:

Let $\left[\boldsymbol{X}_1, \boldsymbol{X}_2 \cdots \boldsymbol{X}_{m-1}, \boldsymbol{X}_M\right]^T$ represent the full-life vibration signal of the M group of rolling bearings, and $\boldsymbol{X}_i = \left[\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2} \cdots \boldsymbol{x}_{i,n-1}, \boldsymbol{x}_{i,N}\right]^T$ be the full-life cycle signal of the *i*-th group, where N is the number of sampling times of the bearing. Therefore, the whole life vibration signals of the group of rolling bearings can generate $\sum_{i=1}^{M} N_i$ group of samples. As shown in Figure 1, samples are input into the constructed one-dimensional convolutional neural network (1DCNN) in batches to perform convolution normalization and other operations. Finally, a neuron is connected to the output end to represent the current service status of the rolling bearing. In this way, the collected samples are sequentially input into the constructed one-dimensional convolutional neural network to obtain the degradation index that characterizes the degradation trajectory of the rolling bearing.

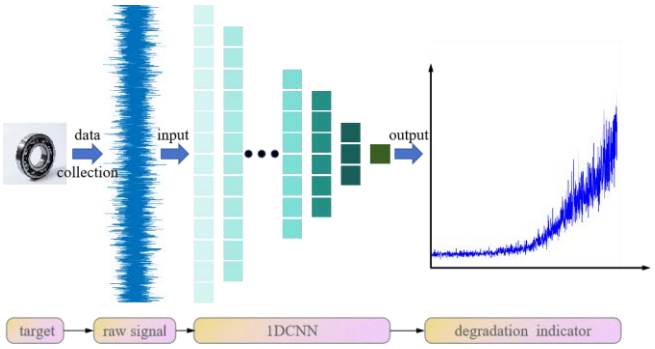


Figure1. DI construction process

2.2. State Assessment Method

The Wiener process has good statistical properties. Therefore, this paper establishes a degradation model of rolling bearings based on the Wiener process to describe its degradation state(Ta et al. 2023). The degradation process of the rolling bearing is described based on the Wiener process as shown in Equation (1), where $y(t)$ represents the degradation state of the rolling bearing at time t, and y_0 is the initial state of the rolling bearing. a is the drift coefficient, which represents the difference between similar rolling bearings and obeys the normal distribution $N(\mu_a,\sigma_a^2)$. t^b is the degradation trend term describing the severity of rolling bearing degradation, where b is a fixed coefficient. c is the diffusion coefficient, which represents the degree of fluctuation when the rolling bearing degrades, and $B(t)$ is the standard Brownian motion (BM), which represents the inherent variability of the random degradation process over time. The fluctuation term describes the uncertainty when the rolling bearing degrades and obeys the normal distribution $N(0,c^2t)$.

$$y(t)=y_0+at^b+cB(t) \tag{1}$$

In order to ensure that rolling bearings always operate safely. Therefore, as shown in equation (2), the RUL l_k of the rolling bearing at time k is defined based on the first hitting time(Cheng et al. 2023), where ω is the failure threshold.

$$l_k=\inf\left\{l:y\left(l+t_k\right)\geq\omega\middle|y\left(t_k\right)=y_k\right\} \tag{2}$$

According to the characteristics of BM and the definition of the RUL of the above formula, the probability density function (PDF) of the RUL of the rolling bearing at any time is shown in Equation (3)(Si et al. 2012).

$$\begin{aligned}f\left(l_k\right)\cong&\frac{1}{\sqrt{2\pi l_k^2\left(\sigma_a^2A\left(l_k\right)^2+\sigma^2l_k\right)}}\\&\times\left(w\left(t_k\right)-B\left(l_k\right)\frac{w\left(t_k\right)\sigma_a^2A\left(l_k\right)+\mu_a\sigma^2l_k}{\sigma_a^2A\left(l_k\right)^2+\sigma^2l_k}\right)\\&\times\exp\left[-\frac{\left(w\left(t_k\right)-\mu_aA\left(l_k\right)\right)^2}{2\left(\sigma_a^2A\left(l_k\right)^2+\sigma^2l_k\right)}\right]\end{aligned} \tag{3}$$

where $A\left(l_k\right)=\left(t_k+l_k\right)^b-t_k{}^b$, $B\left(l_k\right)=A\left(l_k\right)-l_kb\left(t_k+l_k\right)^{b-1}$ and $w\left(t_k\right)=w-y\left(t_k\right)$. After obtaining the PDF of the RUL. As shown in equation (4), the pseudo life is first integrated and averaged, and then the RUL of the rolling bearing at time is obtained(Hu et al. 2020). Then use equation (5) to evaluate the service status of the bearing at the current moment, T_{past} represents the length of time the bearing has been in service relative to the current moment, BC_k represents the service status of the bearing at the current moment, and the closer BC_k is to 100%, the healthier the bearing is.

$$L_k=\int_0^\infty l_kf\left(l_k\right)dl_k \tag{4}$$

$$BC_k=\frac{L_k}{T_{past}+L_k}*100\% \tag{5}$$

According to formula (3) and (4), if the RUL of the rolling bearing at the current time is obtained, the values of parameters μ_a,σ_a^2,b and c^2 need to be estimated. The parameters μ_a,b,c^2 can be obtained using the mapping model (1) as the fitting function. The parameter σ_a^2 can be obtained by the maximum likelihood estimation method. According to the nature of Wiener process, sample $\boldsymbol{y}_{1:N}=\{y_1,y_2,\cdots y_N\}$ follows multivariate normal distribution, let $\boldsymbol{\Lambda}=\left[t_1^b,t_2^b,\cdots t_N^b\right]^T$, then its mean and variance are shown in equation (6):

$$\begin{aligned}&\boldsymbol{y}\sim N\left(\mu_a\boldsymbol{\Lambda},\sigma_a^2\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T+c^2\mathbf{Q}\right)\\&\mathbf{Q}=\left[\min\left\{t_i,t_j\right\}\right]_{1\leq i,j\leq N}\end{aligned} \tag{6}$$

Obtain the PDF of the multivariate normal distribution according to Equation (6) and take the logarithm of both sides to obtain the likelihood function containing unknown parameters. Then use the likelihood function to partially derive the parameter σ_a^2, and make the equation equal to 0.

The solution expression for parameter σ_a^2 is obtained as shown in Equation (7):

$$\sigma_a^2 = \frac{(y_{1:M} - \mu_a \mathbf{\Lambda})^T \mathbf{Q}^{-1} \mathbf{\Lambda} \mathbf{\Lambda}^T \mathbf{Q}^{-1} (y_{1:M} - \mu_a \mathbf{\Lambda}) - c^2 \mathbf{\Lambda}^T \mathbf{Q}^{-1} \mathbf{\Lambda}}{\left(\mathbf{\Lambda}^T \mathbf{Q}^{-1} \mathbf{\Lambda}\right)^2} \quad (7)$$

2.3. Method framework

The proposed method is shown in Figure 2. This method first divides the obtained full-life data into $\sum_{i=1}^{M} N_i$ samples according to the number of collections, and performs data processing on each sample to remove abnormal points and avoid interference with the DI construction model. Secondly, input the processed data into the constructed 1DCNN in batches to train the network until the network converges. Then, the trained network is used as the DI extractor of the rolling bearing, and the newly collected data is input into the DI extractor in sequence according to the number of sampling times, so as to obtain the DI describing the historical operating status of the rolling bearing. Then, use the historical DI data of the rolling bearing to estimate the unknown parameters in the mapping model, and bring them into equations (3) and (4) to obtain the RUL of the rolling bearing at the current moment. Finally, equation (5) is used to evaluate the current service status of the rolling bearing.

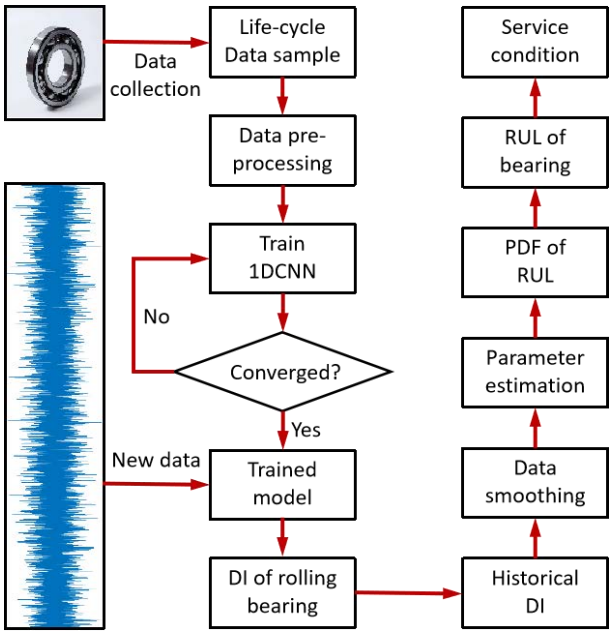


Figure2. Method framework

3. EXPERIMENT

In order to verify the effectiveness of the method, this paper uses two sets of public full-life rolling bearing data sets for verification. The constructed DI is quantitatively analyzed using robustness (Rob), monotonicity (Mon), trendability (Tre) and comprehensive evaluation methods (Com)(Ta et al. 2023). If these four evaluation indicators are larger, it means that the constructed DI can better characterize the degradation trajectory of the bearing. Similarly, in order to analyze the prediction results from a quantitative perspective, this paper uses root mean square error (RMSE), adaptability (R^2), mean absolute error (MAE) and cumulative relative accuracy (CAR) to analyze the prediction results. The smaller the RMSE and the MAE, the better the prediction effect; the larger R^2 means the model has stronger adaptability; the greater the CAR, the better the prediction effect.

3.1. Case 1

Case 1 uses the full-life bearing data provided by the IEEE PHM 2012 Challenge to verify the method. Experimental data comes from PRONOSTIA experimental bench. This data set contains a total of 17 sets of accelerated degradation experimental data of rolling bearings, which were completed under three working conditions, as shown in Table 1. The operating conditions of the 17 sets of rolling bearings are shown in Table 2.

Table1. Operating conditions table

Condition number	Conditions 1	Conditions 2	Conditions 3
Rotating speed	1800 rpm	1650 rpm	1500 rpm
Apply load	4000 N	4200 N	5000

Table2. IEEE PHM 2012 Dataset

Data set	Conditions 1	Conditions 2	Conditions 3
Training set	Bearing1_1	Bearing2_1	Bearing3_1
	Bearing1_2	Bearing2_2	Bearing3_2
Test set	Bearing1_3	Bearing2_3	Bearing3_3
	Bearing1_4	Bearing2_4	
	Bearing1_5	Bearing2_5	
	Bearing1_6	Bearing2_6	
	Bearing1_7	Bearing2_7	

In this experiment, each group of bearings used two vibration sensors to collect data. The sampling frequency was 25.6kHZ, the sampling interval was 10 seconds, and the duration of each sampling was 1 second. In this experiment, this paper uses Bearing1_3 as a test sample, and the others as training samples to train the network, and continuously adjust the network parameters until the network converges. Bearing1_3 data samples are input into the DI extractor successively, and the output DI are smoothed successively. The DI of Bearing1_3 is shown in Figure 3. The constructed DI is compared with the 7 commonly used DI of rolling bearings. The comparison results are shown in Table 3 (Proposed method (M1), Degenerate angle (M2), Maximum value (M3), Mean absolute value (M4), Root mean square (M5) Root

amplitude (M6), Standard deviation (M7) Variance (M8)). It can be seen from the table that the DI constructed using the proposed method has good Tre, Rob and Mon. Because the range of these three evaluation indicators is between [0,1]. Therefore, the three of them are added to form a Com. Judging from the comprehensive indicator column in the table, the DI constructed in this paper is the best.

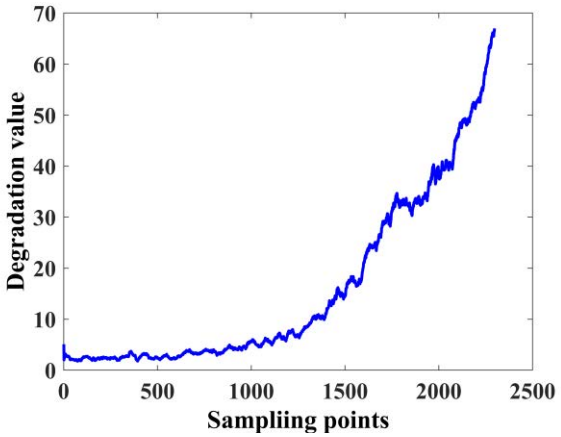


Figure3. Bearing1_3 DI

Table3. Performance comparison of 8 DIs

	Rob	Mon	Tre	Com
M1	0.9932	0.8484	0.8867	2.7283
M2	0.9932	0.1276	0.4102	1.531
M3	0.9797	0.4599	0.7402	2.1798
M4	0.9737	0.3571	0.7402	2.0710
M5	0.7311	0.4207	0.8216	1.9734
M6	0.5934	0.0113	0.2156	0.8203
M7	0.9931	0.4233	0.8145	2.2309
M8	0.9909	0.4382	0.7979	2.2270

Bearing1_3 conducted a total of 2375 samples in the experiment. In order to make the intervals between each condition monitoring (CM) point equal, this paper took the first 2300 samples as test samples, in which the monitoring interval was 100. Finally, Bearing1_3 was monitored 23 times according to the service process of the bearing. The *k*-th CM point represents the service status of the bearing at time *k*, and the previous *k*-th CM point represents the historical service status of the bearing at time *k*. The constructed DIs are input into the PDF of the RUL in batches and the corresponding unknown parameters are estimated. The obtained PDF of the RUL is shown in Figure 4. It can be seen from the figure that with more and more historical data, the PDF becomes more and more convergent, indicating that the credibility of the prediction is getting higher and higher.

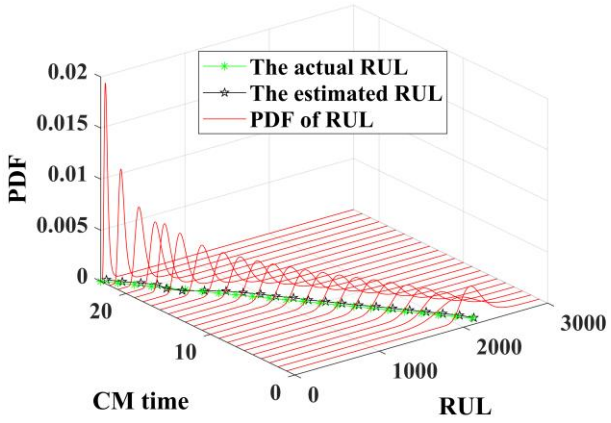


Figure4. PDF of RUL

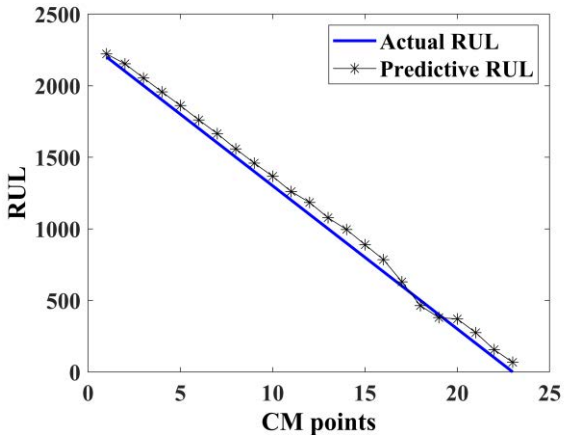


Figure5. Prediction results at different CM points

Table4. Quantitative analysis of prediction results

RMSE	R^2	MAE	CAR
63.5707 （2.76%）	0.9908	60.5620 （2.63%）	0.8610

As can be seen from Figure 5, the prediction results of different CM points are close to the actual RUL of the rolling bearing Bearing1_3. It can be seen from Table 4, the RMSE of the prediction result is only 2.76%, the MAE is 2.63%, R2 is close to 1, and the CAR is 86.10%. The above analysis results show that the method has good accuracy. In addition, Figure 6 shows the service status of the rolling bearing Bearing1_3 at different CM points. It can be seen from the figure that the service performance of the rolling bearing

Bearing1_3 gradually decreases as its service time becomes longer. It also illustrates the effectiveness of this method for evaluating the service status of rolling bearings.

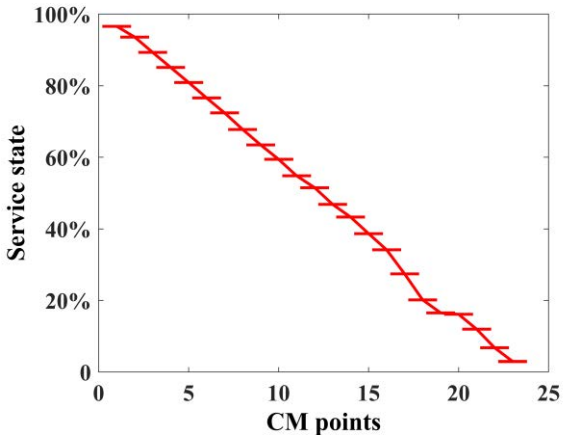


Figure6. Bearing1_3 service status

3.2. Case 2

Case 2 uses the public data set of XJTU-SY for verification. This data set contains a total of 15 sets of full-life bearing data. The sampling frequency is 25.6 kHz, the sampling interval is 1min, and each sampling is 1.28 seconds long. In the same verification method as Case 1, 14 sets of bearings are used as training samples and 1 set is used as test samples. The test sample is Bearing 3_1. The DI of Bearing 3_1 obtained after the final test is shown in Figure 7. It can be seen from the figure that although the DI produces large local volatility, the overall Tre and Rob show good performance. In addition, the performance comparison of different DIs in Table 5 also proves that the DI constructed by this method has good representation performance.

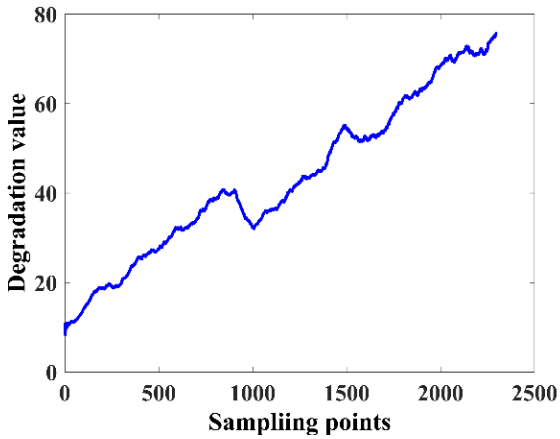


Figure7. Bearing1_3 DI

Table5. Performance comparison of 8 DIs

	Rob	Mon	Tre	Com
M1	0.9431	0.7887	0.9530	2.6848
M2	0.9968	0.0292	0.1321	1.1581
M3	0.9919	0.0252	0.3321	1.3492
M4	0.9919	0.0996	0.3423	1.4338
M5	0.6331	0.0548	0.0941	0.7820
M6	0.5731	0.0236	0.1641	0.7608
M7	0.9960	0.1204	0.3419	1.4583
M8	0.9961	0.1064	0.3427	1.4452

Bearing 3_1 took a total of 2538 samples. In order to keep the monitoring interval unchanged, the first 2500 sampling points were taken for verification, and a total of 25 times of monitoring were conducted. The RUL of PDF for each monitoring is shown in Figure 8. It can be seen from the figure that with more and more historical data, the PDF becomes more and more convergent, which shows that the credibility of the prediction is getting higher and higher. This leads to the same conclusion as Case 1.

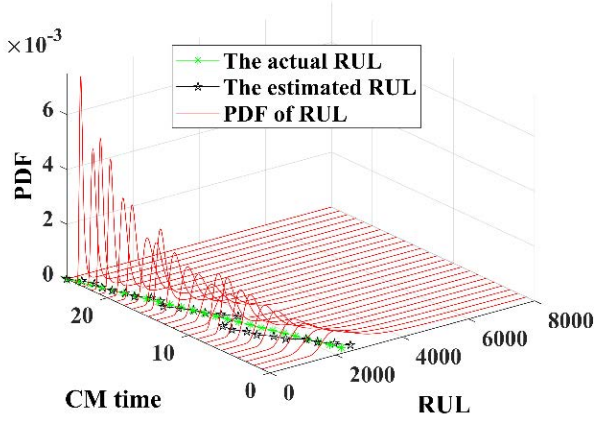


Figure8. PDF of RUL

Figure 9 shows the results of the actual life and predicted life of Bearing 3_1 at different monitoring points. It can be seen from the figure that the deviation at some CM points is larger, and the deviation at some CM points is smaller. This is because the constructed DI has greater volatility at this CM point, which in turn leads to a greater deviation between the prediction results and the actual results. However, from the overall prediction effect, the prediction results are gradually closer to the actual prediction results. As can be seen from

Table 6, the RMSE of the prediction result is 9.72%, MAE is 7.10%, R2 is 0.8865, and CAR is 77.01%. The above analysis results show that the method has good accuracy. Figure 10 is the result of mapping the predicted RUL to service performance, and then determines the service status of the bearing. It can be seen from the figure that as the service time of the bearing increases, the performance of the bearing gradually decreases. Although there was a "recovery" during the period, this can be considered as the self-healing behavior of the bearing during service. Therefore, this method can well evaluate the service status of bearings.

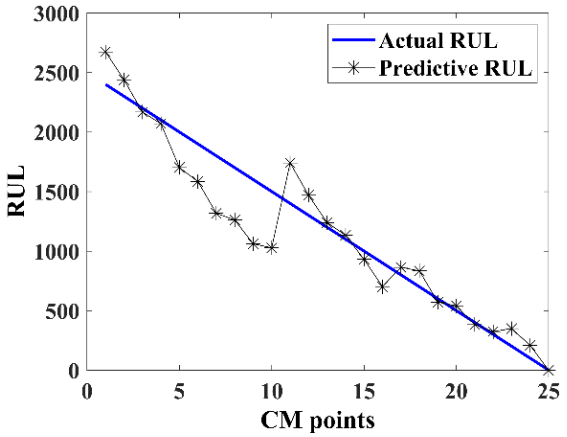


Figure9. Prediction results at different CM points

Table6. Quantitative analysis of prediction results

RMSE	R^2	MAE	CAR
242.9127 （9.72%）	0.8865	177.4692 （7.10%）	0.7701

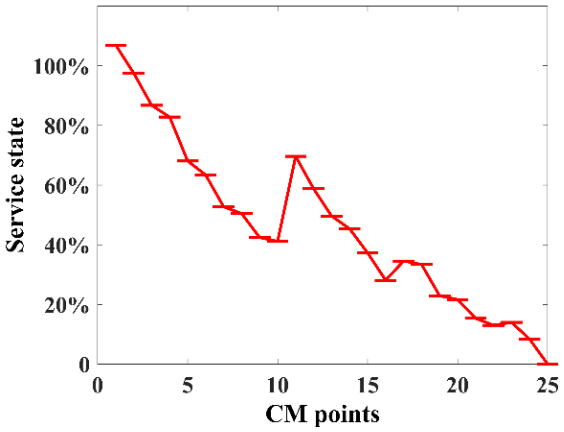


Figure10. Bearing1_3 service status

4. CONCLUSION

In order to evaluate the service status of rolling bearings, this paper proposes a rolling bearing status evaluation method based on deep learning combined with Wiener process. Since the existing DIs cannot characterize the degradation trajectory of rolling bearings. This paper uses a 1DCNN to extract the DIs of rolling bearings. Aiming at the problem of the RUL of rolling bearings, this paper constructs a degradation model of rolling bearings based on the Wiener process, and uses its PDF to estimate the RUL of rolling bearings. The RUL of the rolling bearing is mapped to its service status, thereby completing the service status assessment of the rolling bearing. This paper uses the IEEE PHM 2012 public data set to verify the method. The experimental results show that the extracted DI has good trend and monotonicity, and the service status assessment of the rolling bearing has good accuracy. However, the contribution of this paper is limited. From the verification results, the bearing prediction accuracy is largely determined by the constructed DI and the complexity of the model. Therefore, the follow-up work of this paper will start from mining the degradation information of bearings and establishing more complex prediction models to improve the prediction accuracy of bearings.

REFERENCES

Cheng, W., S. S. Xie, J. Xing, Z. L. Nie, X. F. Chen, Y. L. Liu, X. Liu, Q. Huang, and R. Y. Zhang. 2023. 'Interactive Hybrid Model for Remaining Useful Life Prediction with Uncertainty Quantification of Bearing in Nuclear Circulating Water Pump', *Ieee Transactions on Industrial Informatics*.

Cheng, Y. W., K. Hu, J. Wu, H. P. Zhu, and X. Y. Shao. 2021. 'A convolutional neural network based degradation indicator construction and health prognosis using bidirectional long short-term memory network for rolling bearings', *Advanced Engineering Informatics*, 48.

Hu, C. H., H. Pei, X. S. Si, D. B. Du, Z. N. Pang, and X. Wang. 2020. 'A Prognostic Model Based on DBN and Diffusion Process for Degrading Bearing', *Ieee Transactions on Industrial Electronics*, 67: 8767-77.

Kogan, G., R. Klein, A. Kushnirsky, and J. Bortman. 2015. 'Toward a 3D dynamic model of a faulty duplex ball bearing', *Mechanical Systems and Signal Processing*, 54-55: 243-58.

Li, T. M., X. S. Si, H. Pei, and L. Sun. 2022. 'Data-model interactive prognosis for multi-sensor monitored stochastic degrading devices', *Mechanical Systems and Signal Processing*, 167.

Li, Y. J., Z. J. Wang, F. Li, Y. F. Li, X. H. Zhang, H. Shi, L. Dong, and W. B. Ren. 2024. 'An ensembled remaining useful life prediction method with data fusion and stage division', *Reliability Engineering & System Safety*, 242.

Qian, Y. N., R. Q. Yan, and R. X. Gao. 2017. 'A multi-time scale approach to remaining useful life prediction in rolling bearing', *Mechanical Systems and Signal Processing*, 83: 549-67.

Qin, Y., J. H. Yang, J. H. Zhou, H. Y. Pu, and Y. F. Mao. 2023. 'A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery RUL prediction', *Advanced Engineering Informatics*, 56.

Ren, L., Y. Q. Sun, J. Cui, and L. Zhang. 2018. 'Bearing remaining useful life prediction based on deep autoencoder and deep neural networks', *Journal of Manufacturing Systems*, 48: 71-77.

Rezamand, M., M. Kordestani, M. E. Orchard, R. Carriveau, D. S. K. Ting, and M. Saif. 2021. 'Improved Remaining Useful Life Estimation of Wind Turbine Drivetrain Bearings Under Varying Operating Conditions', *Ieee Transactions on Industrial Informatics*, 17: 1742-52.

She, D. M., and M. P. Jia. 2019. 'Wear indicator construction of rolling bearings based on multi-channel deep convolutional neural network with exponentially decaying learning rate', *Measurement*, 135: 368-75.

Si, X. S., W. B. Wang, C. H. Hu, D. H. Zhou, and M. G. Pecht. 2012. 'Remaining Useful Life Estimation Based on a Nonlinear Diffusion Degradation Process', *Ieee Transactions on Reliability*, 61: 50-67.

Ta, Y. T., Y. F. Li, W. A. Cai, Q. Q. Zhang, Z. J. Wang, L. Dong, and W. H. Du. 2023. 'Adaptive staged remaining useful life prediction method based on multi-sensor and multi-feature fusion', *Reliability Engineering & System Safety*, 231.

Wang, B., Y. G. Lei, N. P. Li, and N. B. Li. 2020. 'A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings', *Ieee Transactions on Reliability*, 69: 401-12.

Wang, X., L. L. Cui, and H. Q. Wang. 2022. 'Remaining Useful Life Prediction of Rolling Element Bearings Based on Hybrid Drive of Data and Model', *Ieee Sensors Journal*, 22: 16985-93.

Wang, Z. J., Y. T. Ta, W. N. Cai, and Y. F. Li. 2023. 'Research on a remaining useful life prediction method for degradation angle identification two-stage degradation process', *Mechanical Systems and Signal Processing*, 184.

Xiahou, T. F., Z. G. Zeng, and Y. Liu. 2021. 'Remaining Useful Life Prediction by Fusing Expert Knowledge and Condition Monitoring Information', *Ieee Transactions on Industrial Informatics*, 17: 2653-63.

Yoo, Y., and J. G. Baek. 2018. 'A Novel Image Feature for the Remaining Useful Lifetime Prediction of Bearings Based on Continuous Wavelet Transform and Convolutional Neural Network', *Applied Sciences-Basel*, 8.

Zhu, D., J. W. Lyu, Q. W. Gao, Y. X. Lu, and D. W. Zhao. 2024. 'Remaining useful life estimation of bearing using spatio-temporal convolutional transformer', *Measurement Science and Technology*, 35.

A Semi-supervised Fault Diagnosis Method Based on Graph Convolution for Few-shot Fault Diagnosis

Yuyan Li[1], Tiantian Wang[2], and Jingsong Xie[3]

[1,3] *College of Traffic and Transportation Engineering, Central South University, Changsha, China*

liyuyan@csu.edu.cnu

jingsongxie@csu.edu.cn

[2] *College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China*

wangtt@hnu.edu.cn

ABSTRACT

In practical bearing fault diagnosis, labeled fault data are difficult to obtain, and limited samples will lead to training overfitting. To address the above problems, a semi-supervised fault diagnosis method based on graph convolution is proposed. Firstly, the KNN graph construction method based on Euclidean distance (ED-KNN) is used to achieve label propagation. Then, a graph convolutional network framework based on dot product attention mechanism (GPGAT) was constructed to enhance the weights of high similarity nodes and diagnose bearing faults. The proposed method is validated on a public bearing dataset. The results show that the proposed method can make full use of very few labeled samples for fault diagnosis. Compared with other state-of-the-art methods, the proposed method achieves better diagnosis performance.

1. INTRODUCTION

Rotating machinery plays a crucial role in manufacturing, industrial robotics, transportation, and other fields. Bearings, as vital components of rotating machinery, may lead to significant economic losses if they fail. Bearings generate vast amounts of data during operation, and how to extract useful information from this data has become a hot topic in bearing fault diagnosis research in recent years n (Zhang et al., 2023). Intelligent fault diagnosis is an automated reasoning process based on data-driven approaches. In recent years, various deep learning models have been successfully applied to intelligent fault diagnosis (Jiao et al., 2020). However, their effective training relies on a large amount of labeled data, which is quite challenging in practical fault diagnosis (Yang et al., 2023). In engineering, labeling and screening data are time-consuming tasks, making it essential to study high-precision bearing fault diagnosis methods under extremely scarce labeled samples.

Semi-supervised learning can leverage a small number of labeled samples to learn the information contained in the vast majority of unlabeled samples. In recent years, it has been widely studied in intelligent fault diagnosis in mechanical systems. Ding et al. (2023) trained multiple GANs to eliminate abnormal cases, thereby enhancing the performance of small-sample fault diagnosis in a semi-supervised manner. Yu et al. (2020) investigated a data augmentation method based on consistency regularization, which achieved fault diagnosis of bearings in cases where labeled samples are limited. Zhang et al. (2019) proposed an Active Semi-Supervised Learning GAN (ASSL-GAN), which minimizes the loss function through alternate updates to achieve higher accuracy. These methods can to some extent address the challenge of insufficient labeled samples in fault diagnosis tasks.

In recent years, with the flourishing development of Graph Neural Networks (GNNs) (Scarselli, F. et al., 2019), graph-based semi-supervised algorithms have gradually become a research hotspot. A graph based semi supervised learning algorithm propagates labeled data labels to unlabeled data by constructing a graph. The following paper provides a similar method implementation. Xie et al. (2022) utilized multi-scale graph convolution to aggregate multi-scale information of labeled samples and introduced an attention mechanism to form a new adaptive feature fusion layer. They proposed the Semi-supervised Multi-Scale Attention Graph Convolutional Network (MSA-GCN) for fault diagnosis and achieved satisfactory results. Kavianpour et al. (2022) addressed the issues of insufficient labeling of fault diagnosis data, changing operating conditions, and data loss in practical applications by aligning subdomains of the same class. They proposed a semi-supervised method based

on Autoregressive Moving Average (ARMA) filter graph convolution, adversarial adaptation, and Multi-layer Multi-kernel Local Maximum Mean Discrepancy (MK-LMMD). The above literature demonstrates the unique advantages of Graph Neural Networks in semi-supervised learning. However, this research still faces challenges such as high labeling rates and model instability under extreme labeling conditions, and there are still some shortcomings in feature mining for low-labeled samples, limiting its practical application.

In response to the scenario of fault diagnosis with extremely few labeled data, this paper proposes a network that combines Euclidean distance-based KNN graph (ED-KNN) with dot product graph attention mechanism (DPGAT). By utilizing Euclidean distance to measure the distance between labeled and unlabeled samples, an accurate KNN graph is obtained. Then, the dot product attention mechanism is used to further increase the weights of neighboring nodes with high similarity, in order to learn the optimal representation of the graph. The proposed method is experimented on a publicly available bearing dataset. The results demonstrate that the proposed method achieves high-precision classification of unlabeled data with minimal training on extremely few samples, indicating its significant engineering application value.

2. RELATED THEORIES

2.1. Graph neural networks

Unlike convolutional neural networks (CNNs), graph neural networks (GNNs) are a class of learning models based on graph-structured data. They can define graph convolutions based on the connections between nodes in non-Euclidean space. The difference between CNNs and GNNs in terms of convolution can be intuitively illustrated as shown in Figure 1. The involved graph structure can be simplified as follows:

$$G = (H, A) \tag{1}$$

Here, $H = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\} \in R^{n \times d}$ represents the set of nodes; n is the number of nodes; d is the dimensionality of the input node features; $A \in R^{n \times n}$ represents the adjacency matrix representing the connections between nodes. The graph convolutional layer updates node features by aggregating neighboring node features. Typically, given the input graph G, the convolutional layer outputs a new set of node features $\boldsymbol{H}' = \{\boldsymbol{h}_1', \boldsymbol{h}_2', \cdots, \boldsymbol{h}_n'\} \in R^{n \times d'}$ with dimension d'. The graph convolutional layer can be represented as:

$$h_i^{'} = \Gamma(h_i, \mathrm{Y}(\{\boldsymbol{h}_j \mid j \in N_i\})) \tag{2}$$

Among them, N_i is the number of neighboring nodes of node hi; $\Gamma(\cdot)$ represents nonlinear layers; $\mathrm{Y}(\cdot)$ represents a certain node aggregation pattern.

2.2. Semi-supervised Learning with GNN

Graph-based semi-supervised learning typically involves establishing explicit relationships between labeled data and a large amount of unlabeled data using a graph structure, where data points are represented as vertices and the similarity between points is represented as edges. The constructed graph is then inputted into a graph neural network to obtain feature-level representations of the graph and its nodes. These graph-level or node-level features are then fed into a classifier for classification and fault diagnosis. This process leverages the graph structure to effectively utilize both labeled and unlabeled data for semi-supervised fault diagnosis.

3. PROPOSED METHOD

3.1. ED-KNN

Graphs can represent the similarity relationships between samples. Initially, the time-domain vibration signals collected from bearings are segmented via multiple sampling. Subsequently, these segments are transformed into frequency-domain signals using Fast Fourier Transform (FFT). The KNN graph is constructed by assessing the adjacency between labeled and unlabeled samples using Euclidean distance. The distance metric formula utilizing Euclidean distance is:

$$dis(x_i, y_i) = \left(\sum_{l=1}^{d} |x_i - y_i|^2 \right)^{\frac{1}{2}} \tag{3}$$

Where x_i represents the feature of the central node and y_i represents the neighboring node of x_i. For a certain central node $x' \in x_i$, the distance values between it and other neighboring nodes are arranged in ascending order:

$$D = \{dis_1(x', y_1), ..., dis_n(x', y_n)\}, (dis_1 < ... < dis_n) \tag{4}$$

The neighboring nodes of node x' are selected through k-nearest neighbors, denoted as:

$$\mathrm{Top-k} = \{x_1', x_2', ..., x_k'\} \tag{5}$$

Top-k represents the set of k-nearest neighbors of x_i, where k is the number of nearest neighbors. Through experiments, it has been found that when k is set to 5, the quality of the constructed graph is satisfactory. By constructing the ED-KNN graph, each time k unlabeled data points are assigned pseudo-labels. This step establishes an intrinsic graph structure connection between labeled and unlabeled observed data, which can be regarded as a form of label propagation process.

When applying KNN nearest neighbor search, the sample set consists of all the samples from the bearings in that sample set. The connecting nodes are selected based on the proximity

determined by distances, where the top K nearest neighbors are chosen.

ED-KNN graph construction utilizes the similarity of feature vectors among samples to establish connections between them. By exploiting the joint dependencies between labels, label information is propagated along these connections, enabling the assignment of pseudo-labels to unlabeled samples. This approach facilitates a more thorough exploration of limited label information, thereby augmenting the model's capacity to learn from label information. The intuitive workflow is depicted in Fig. 1. Among them, when constructing the training set, the ED-KNN graph is constructed with labeled data as the central node and unlabeled data as neighboring nodes. Due to the large number of unlabeled data, there will be unlabeled data that has not been assigned and will not participate in model training. Meanwhile, the same sample may also be repeatedly labeled and participate in the construction of graphs with different central nodes. It is worth noting that since we use the entire graph for training, Neighboring nodes output features through weighted output. Therefore, nodes that are repeatedly labeled will not affect training, as they will be assigned different node weights in different graphs. When constructing the test set, all unlabeled data points in the test set are sequentially used to calculate the Euclidean distance from all other samples, and the top K-nearest samples are selected as neighboring nodes. Therefore, the number of constructed graphs is the same as the number of samples in the test set. Labeled data is only provided during the training phase, while in the testing phase, there is no availability of labeled data. When performing convolution calculations after constructing a graph, it is necessary to ensure that the number of linked nodes is consistent, otherwise graph convolution calculations will be very difficult. Therefore, we construct a KNN graph based on the top 5 nearest neighboring nodes. In fact, calculating the distance between nodes, selecting nodes through counting, and selecting nodes through threshold are similar.

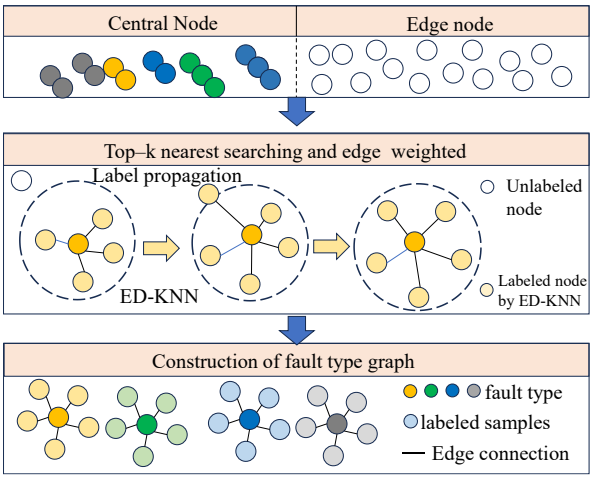


Figure 1. The process of ED-KNN.

3.2. DPGAT Diagnosis Framework

This paper proposes utilizing the dot-product attention mechanism to better learn graph representations by computing the weights of neighboring nodes' influence on the central node. Let $h=\{h_1,h_2,\ldots,h_N\}, h_i \in \mathbb{R}^F$ denote the input features of nodes, where N and F represent the number of nodes and the feature dimension, respectively. The output features of nodes are denoted as $h'=\{h_1',h_2',\ldots,h_N'\}, h_i' \in \mathbb{R}^{F'}$. $W \in \mathbb{R}^{F' \times F}$ represents the weight matrix of linear transformations applied at each node. Finally, softmax normalization is applied, followed by Leaky ReLU to introduce non-linearity. The output features of nodes are obtained using the following equation:

$$\mathbf{h}_i' = \sum_{j \in N_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j \tag{6}$$

α_{ij} signifies the attention coefficient from neighboring node j to central node i, reflecting the significance of node j with respect to node i. α_{ij} is derived through SoftMax normalization of the attention parameter e_{ij} for each edge. The expression for the attention coefficient of node pair (i,j) is given by:

$$\alpha_{ij} = \mathrm{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \tag{7}$$

The attention parameter e_{ij} is obtained through the outer product attention mechanism, which originates from node representation learning (Kim, D., & Oh, A. , 2022). The outer product of a node with its transpose can be regarded as its attention score. The mathematical expression for the outer product attention mechanism is:

$$e_{ij} = (\mathbf{W}h_i)^{\mathrm{T}} \cdot \mathbf{W}h_j \tag{8}$$

Plugging it into Eq. (7) enables us to obtain the attention coefficients for each pair of nodes.

$$\alpha_{ij} = \mathrm{softmax}(e_{ij}) \tag{9}$$

The features outputted by DPGAT are inputted into a fully connected (FC) layer to obtain the predicted label set. The prediction process can be represented as:

$$z_i = FC(\mathbf{h}_i') \tag{10}$$

The loss function of the DPGAT is:

$$loss_{Class} = -\frac{1}{I} \sum_i^I \sum_t^T y_i^{(t)} \ln(z_i^{(t)}) \tag{11}$$

Here, I denotes the label index; T stands for the number of classes; $y_i^{(t)}$ represents the t-dimensional value of the true labels; and $z_i^{(t)}$ signifies the t-dimensional value of the predicted label z_i.

3.3. The Overall Procedure

The overall framework of the proposed method is shown in Fig. 2, with specific explanations as follows.

1) Signal Acquisition: Collect vibration signals from bearings using sensors on the experimental platform.

2) Graph Construction: Divide the collected vibration signals into multiple independent samples and split them into training and testing sets. The training set comprises a small amount of labeled data and real collected data, while the testing set consists only of unlabeled data. Utilize the ED-KNN method to obtain the KNN graph.

3) Model Training: Construct a feature extraction network based on DPGAT. Obtain output features through Eq. (6). Input the training set sequentially into two DP-GAT layers and two FC layers, and obtain the predicted label set through Eq. (10). Then compute the loss using Eq. (11).

4) Model Testing: Feed the unlabeled testing set into the trained model to obtain diagnostic results and compare them with other semi-supervised fault diagnosis methods based on common GNNs.

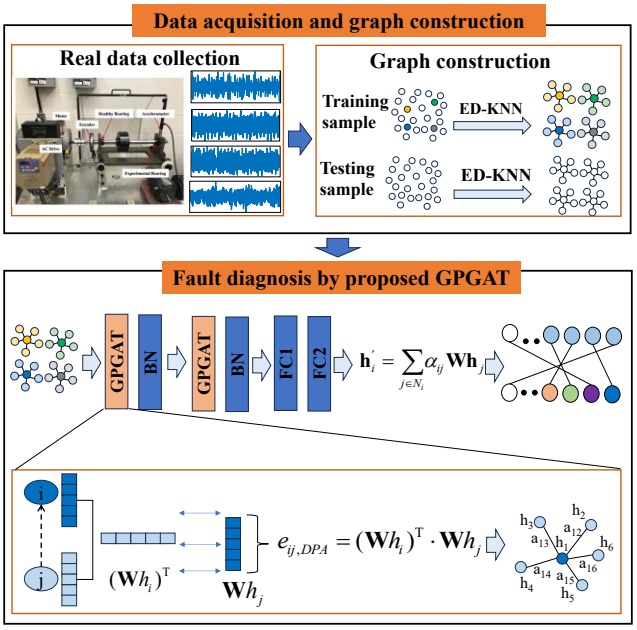


Figure 2. The framework of proposed method.

4. Validation Of Performance

The effectiveness of the proposed method was validated through two semi-supervised fault diagnosis instances. In Case 1, the dataset from Case Western Reserve University (Smith et al., 2015) was utilized, while in Case 2, the dataset from the University of Ottawa (UofO) (Huang & Baddour, 2018) was employed. To demonstrate the superiority of the proposed approach, it was compared with seven widely researched graph neural networks, including Basic GAT (Veličković et al., 2018), DGAT, Graph Transformer (Shi et al., 2021), GraphConv (Morris et al., 2019), ChebConv (Defferrard et al., 2016), GraphSage (Hamilton et al., 2017), and GEN (Li et al., 2016). The above methods are only for graph convolutional models and do not involve a semi supervised learning process. We put it into the semi supervision framework proposed in this paper (using ED-KNN construction diagram) to verify the progressiveness of the proposed GPGAT.

4.1. Case 1: CWRU Dataset

The CWRU dataset was tested using SKF 6205 drive-end bearings. The sampling frequency of the accelerometer was 48 kHz. The bearing loads were categorized as 0HP, 1HP, 2HP, and 3HP, with corresponding speeds of 1797rpm, 1772rpm, 1750rpm, and 1730rpm, respectively. The health conditions of the bearings included four forms: Inner Race Fault (IF), Rolling Element Fault (ReF), Outer Race Fault (OF), and Normal Condition (NC). For each health condition's vibration signal, a sampling length of 1024 and the same sampling interval are used to ensure that there is no repetition between the data, resulting in 400 samples. These 400 samples were then randomly divided into training and testing samples at a ratio of 1:1. Verify the effectiveness of the proposed method through accuracy validation on the test set

Table 1. Description of the CWRU dataset.

Fault type	Speed(rpm)	Labeled samples and labeled rate	Train	Test
OF IF ReF NC	1730	4×1(0.25%)	4× 200	4× 200

4.2. Case 2: UofO Dataset

The dataset originates from the SpectraQuest Mechanical Fault Simulator at the University of Ottawa. Two ER16K ball bearings were installed to support the rotating shaft, which could be replaced with bearings in different health states. Accelerometers (ICP accelerometer, model 623C01) were placed on the experimental bearing housing for vibration data collection, while an incremental encoder (model EPC-775) measured the shaft speed. The signal sampling frequency was 200 kHz, and each experiment lasted for 10 seconds, including both acceleration and deceleration processes. For Case 2, vibration signals from bearings in four different states, including three types of faults and normal condition, were selected. The length of each sample was 4096 sampling points.

Table 2. Description of the UofO dataset.

Fault type	Speed(rpm)	Labeled samples and labeled rate	Train	Test
OF IF ReF NC	846~1428	4×20(5%)	4×200	4×200

4.3. Experimental Results

To validate the superiority of the proposed construction method, the GPGAT was compared with seven others advanced GNN methods, and the average diagnostic accuracy is shown in Tables 3. In Case 1, the proposed GPGAT achieved a classification accuracy of 98.67%, which is 2.5% higher than the other best-performing method DGAT. In Case 2, the proposed GPGAT achieved a classification accuracy of 97.38%, which is 2.71% higher than the other best-performing methods GAT and ChebConv. The GPGAT proposed in this paper achieved better diagnostic accuracy compared to other graph convolution methods on both datasets, validating the effectiveness of the proposed approach.

Table 3. The test accuracy on the Case 1 and Case 2.

Method	Case1	Case2
GPGAT(Proposed)	**98.67%**	**97.38%**
GAT	95.57%	94.87%
DGAT	96.17%	90.38%
Graph Transformer	90.58%	85.00%
GraphConv	94.17%	93.63%
ChebConv	95.14%	94.87%
SAGE	95.83%	87.17%
GEN	91.75%	91.37%

To further demonstrate the diagnostic performance of the proposed method, we visualize the confusion matrix for Case 2, as shown in Fig. 3. Each health state has 200 test samples. The horizontal axis represents the predicted labels, while the vertical axis represents the true labels, where 0-3 denote the four health states OF, IF, ReF, and NC listed in Table 2. It can be observed that for the multi-class classification task, the proposed method GPGAT exhibits the best diagnostic performance.

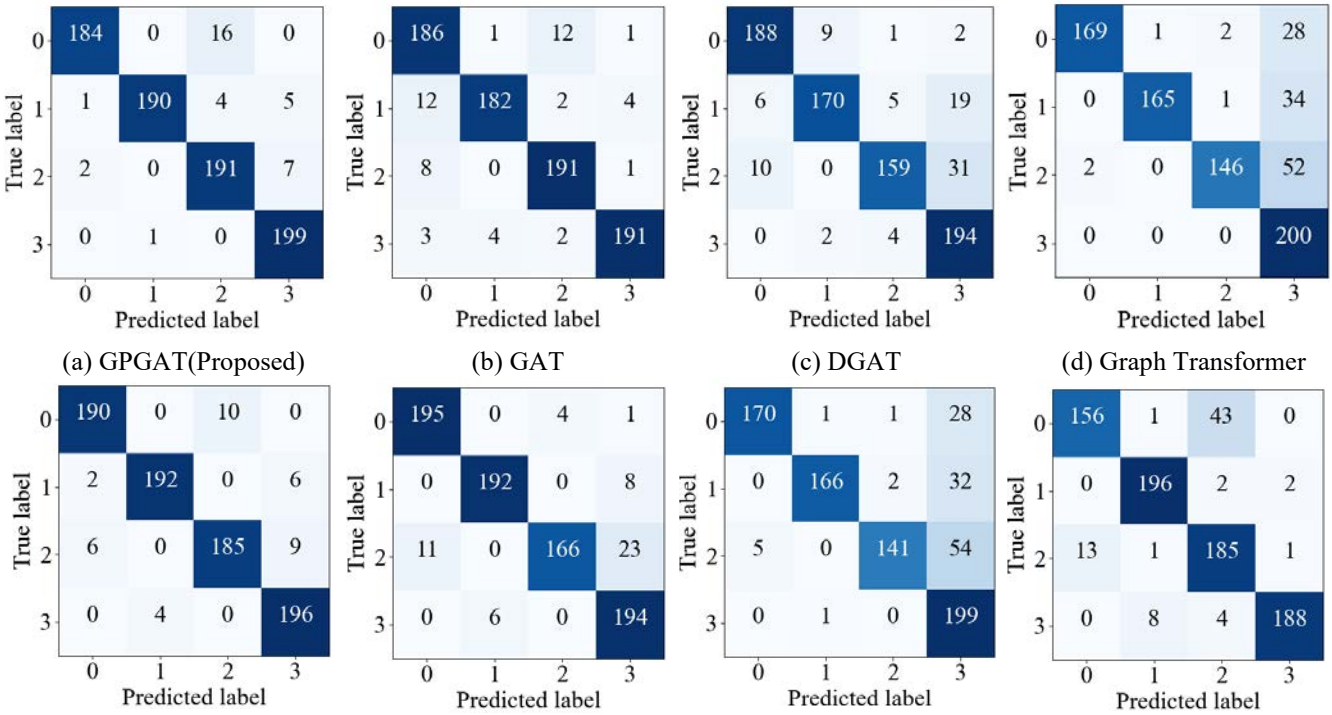
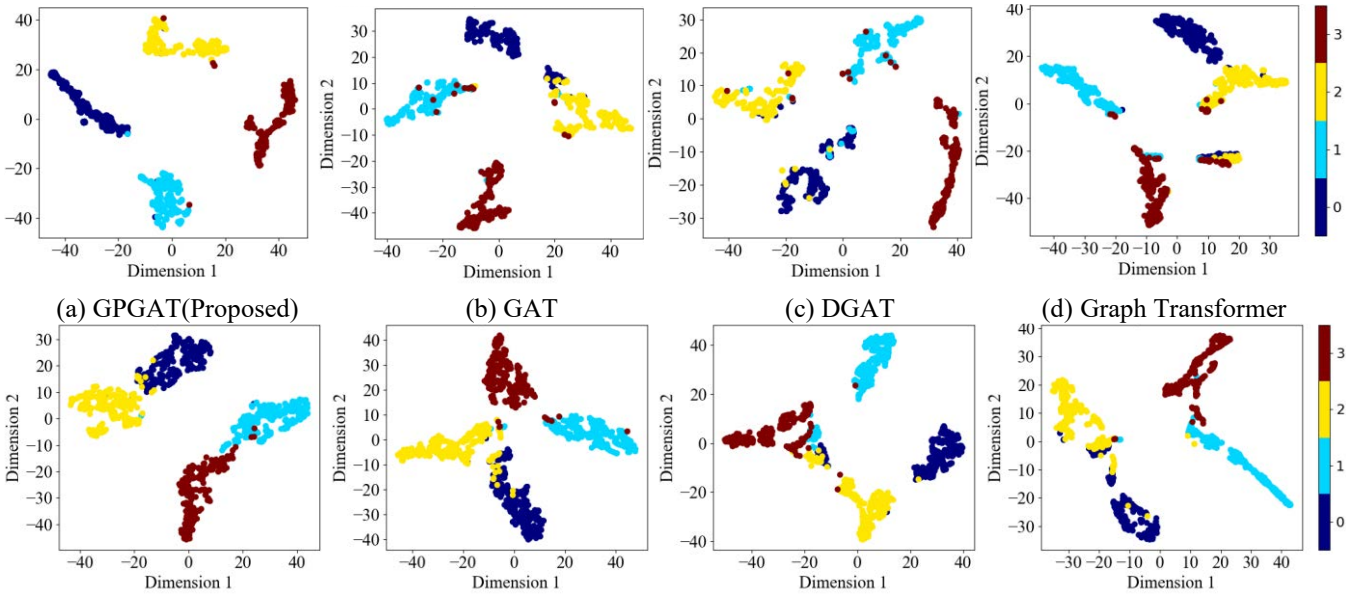


Figure 3. Classification confusion matrix for Case 2.

To better illustrate the feature extraction performance of the proposed method, the output feature vectors are reduced to two dimensions using T-SNE for Case 2, as shown in Fig. 4. From (a), it can be seen that the four types of features

represented by the four colors have a good degree of aggregation, and the distance between each type is relatively far, indicating that the proposed GPGAT has a good feature extraction ability. The features extracted by GPGAT exhibit higher aggregation and greater distances from each other compared to other methods. The proposed method demonstrates better discriminative ability for all health states, and GPGAT maintains good diagnostic performance even at extremely low label rates.



(a) GPGAT(Proposed) (b) GAT (c) DGAT (d) Graph Transformer

(e) ChebConv (f) GraphConv (g) SAGE (h) GEN

Figure 4. 2D visualization of the output features for all the methods on Case 2.

4.4. Discussion on Labeled Rate and K-value

To validate the effectiveness of the proposed method under small sample conditions, Case 2 dataset was trained and tested with labeled samples of 4, 8, 12, 16, and 20. The experiments were repeated ten times to obtain diagnostic accuracy. As shown in Fig. 5, it can be observed that as the label rate increases from 1% to 5%, the testing accuracy continues to improve. Even with a label rate of 1%, a fault diagnosis rate of 93.75% can be achieved, while a label rate of 5% yields a fault diagnosis accuracy of 97.38%.

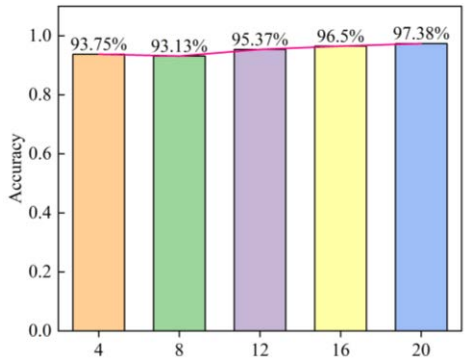


Figure 5. Accuracy of the proposed method with different low labeling rates.

Next, we discuss the influence of the *k* value used to construct the ED-KNN. The *k* value represents the number of neighboring nodes connected to each central node when creating the KNN graph. Using the proposed method, experiments were conducted on the two datasets at a label rate of 1%. Fig. 6 shows the effect of different *k* values on the diagnostic accuracy of the proposed model. It can be observed that on both the CWRU and UofO datasets, the diagnostic accuracy reaches its highest value when *k*=5, with accuracies of 100% and 93.75%, respectively. On the CWRU dataset, the change in accuracy with increasing K values is not particularly significant. This is mainly because the CWRU dataset is collected under steady-state conditions with artificially injected bearing faults, resulting in clean data with very distinct fault characteristics. Therefore, even with smaller K values, a good label propagation efficiency can be maintained. In contrast, the UofO dataset is collected under time-varying conditions, with faults occurring naturally, making fault characteristics less pronounced. Hence, an appropriate K value is required for ED-KNN. A suitable K value ensures that as much unlabeled data as possible is incorporated into the graph while minimizing graph construction errors.

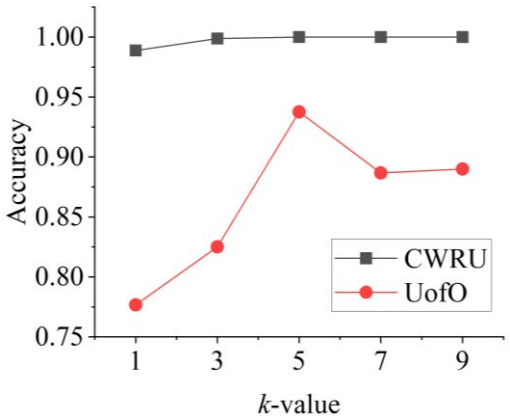


Figure 6. Diagnostic accuracy with different k values.

5. Conclusion

In response to the challenges faced by fault diagnosis under conditions of few samples, this paper proposes a new semi-supervised fault diagnosis method. The proposed ED-KNN calculates the Euclidean distance and sorts the distances in order to obtain the nearest neighboring nodes. It achieves label propagation from labeled data to unlabeled data. The designed GPGAT assigns different importance information to neighboring nodes through the dot product attention mechanism, further enhancing the reliability of the graph. Experimental validation was conducted on the CWRU and UofO dataset. Comparative results indicate that: (1) ED-KNN can effectively construct an undirected graph of labeled and unlabeled data, achieving label propagation. (2) The constructed GPGAT can assign different importance to neighboring nodes, thereby more accurately extracting node features and classification information from the KNN graph. (3) Compared with other state-of-the-art methods, the proposed approach can more accurately diagnose unlabeled samples under conditions of few or even extremely few samples.

References

Zhang, S., Su, L., Gu, J., Li, K., Zhou, L., & Pecht, M. (2023). Rotating machinery fault detection and diagnosis based on deep domain adaptation: A survey. *Chinese Journal of Aeronautics*, 36(1), 45–74. doi: 10.1016/j.cja.2021.10.006

Jiao, J., Zhao, M., Lin, J., & Liang, K. (2020). A comprehensive review on convolutional neural network in machine fault diagnosis. *Neurocomputing*, 417, 36–63. doi: 10.1016/j.neucom.2020.07.088

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9), 8934–8954. doi: 10.1109/TKDE.2022.3220219

Ding, Y., Ma, L., Ma, J., Wang, C., & Lu, C. (2019). A Generative Adversarial Network-Based Intelligent Fault Diagnosis Method for Rotating Machinery Under Small Sample Size Conditions. *IEEE Access*, 7, 149736–149749. doi: 10.1109/ACCESS.2019.2947194

Yu, K., Ma, H., Lin, T. R., & Li, X. (2020). A consistency regularization based semi-supervised learning approach for intelligent fault diagnosis of rolling bearing. *Measurement*, 165, 107987. doi: 10.1016/j.measurement.2020.107987

Zhang, X.-Y., Shi, H., Zhu, X., & Li, P. (2019). Active semi-supervised learning based on self-expressive correlation with generative adversarial networks. *Neurocomputing*, 345, 103–113. doi: 10.1016/j.neucom.2019.01.083

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1), 61–80. doi: 10.1109/TNN.2008.2005605

Xie, Z., Chen, J., Feng, Y., & He, S. (2022). Semi-supervised multi-scale attention-aware graph convolution network for intelligent fault diagnosis of machine under extremely-limited labeled samples. *Journal of Manufacturing Systems*, 64, 561–577. doi: 10.1016/j.jmsy.2022.08.007

Kavianpour, M., Ramezani, A., & Beheshti, M. T. H. (2022). A class alignment method based on graph convolution neural network for bearing fault diagnosis in presence of missing data and changing working conditions. *Measurement*, 199, 111536. doi: 10.1016/j.measurement.2022.111536

Kim, D., & Oh, A. (2022). How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. arXiv. Retrieved from http://arxiv.org/abs/2204.04879

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, 64–65, 100–131. doi: 10.1016/j.ymssp.2015.04.021

Huang, H., & Baddour, N. (2018). Bearing vibration data collected under time-varying rotational speed conditions. *Data in Brief*, 21, 1745–1749. doi: 10.1016/j.dib.2018.11.019

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, February 4). Graph Attention Networks. arXiv. Retrieved from http://arxiv.org/abs/1710.10903

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2021, May 9). Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. arXiv. Retrieved from http://arxiv.org/abs/2009.03509

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., & Grohe, M. (2019). Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4602–4609. doi: 10.1609/aaai.v33i01.33014602

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in Neural Information Processing Systems*, 29. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, 30. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

Li, G., Xiong, C., Thabet, A., & Ghanem, B. (2020, June 13). DeeperGCN: All You Need to Train Deeper GCNs. arXiv. doi: 10.48550/arXiv.2006.07739

BIOGRAPHIES

Yuyan Li received his M.Sc. degree from Kunming University of Science and Technology, Kunming, China. He is currently pursuing his D.E. degree in Traffic and Transportation Engineering at Central South University, Changsha, China. His current research interests include fault diagnosis, machine learning, vibration analysis, and pattern recognition.

Tiantian Wang received his bachelor's and Ph.D. degrees from Beihang University, Beijing, China, in 2012 and 2018, respectively. He is currently a Vice Professor at Central South University and Hunan University. His current research interests include vehicle aerodynamics and vehicle structures, especially train/tunnel aerodynamics and PHM for trains.

Jingsong Xie was born in Anren, Hunan, China, in 1989. He received his B.S. degree from the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China, in 2013, and his Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, in 2018.
He joined the School of Traffic and Transportation Engineering, Central South University, Changsha, China, as a lecturer. His research interests include fault diagnosis, machine learning, vibration analysis, and crack diagnosis.

A Study on the Equipment Data Collection and Developing Next Generation Integrated PHM System

Deog Hyeon Kim[1], Gun Sik Kim[1], Jung Ho Nam[1] and Jin Woo Park[1]

[1]*Hyundai Motor Company, Equipment Control Engineering Team, Ulsan, 44259, South Korea*

dhkims@hyundai.com
6505602@hyundai.com
ionpower@hyundai.com
jin4417@hyundai.com

ABSTRACT

This research presents an integrated PHM system for 2,000 rotating equipment units across press, car body, paint, and assembly lines in Hyundai/Kia factories. The system addresses limitations of individual monitoring systems by consolidating vibration, current, robot AI diagnostics, PLC backup status, and operational data. Vibration monitoring utilizes wired/wireless sensors, server storage, and automated analysis for trend detection and fault diagnosis. PLC data monitoring retrieves motor drive information (current, temperature, frequency, etc.) to predict equipment anomalies. Robot monitoring integrates with various manufacturers and tracks operational status, motor load, and alarms for maintenance and lifespan management. The PLC backup solution ensures proper backup functionality. The integrated PHM architecture manages data collection, analysis, diagnostics, reporting, and visualization, enabling comprehensive equipment health monitoring and proactive maintenance.

1. INTRODUCTION

The optimal approach to equipment maintenance in the factory involves a maintenance strategy divided into reactive, preventive, and predictive methods. Among these, predictive maintenance stands out as an effective way to anticipate failures through equipment condition monitoring [Paulina Gackowiec]. It provides timely insights into breakdown causes, which is increasingly vital due to the industrial internet of things. The shift from reactive to predictive maintenance represents an innovative process improvement. In the reactive maintenance method, urgent repairs must be carried out post-failure, degrading maintenance quality and endangering workers. Conversely, predictive maintenance enables preemptive action, preventing factory shutdowns. By monitoring conditions and analyzing root causes in advance, maintenance can be performed proactively, and equipment condition can be evaluated thereafter. Sudhanshu Goel's paper highlights the significant potential of condition monitoring in enhancing operational reliability, machine uptime, damage reduction, and operational efficiency at a lower cost [Sudhanshu Goel]. Equipment incipient faults often exhibit variations in temperature, vibro-acoustic signature, etc. Different condition monitoring techniques utilize dedicated sensing and data analysis tools to analyze specific operational characteristic variations [Figure 1].

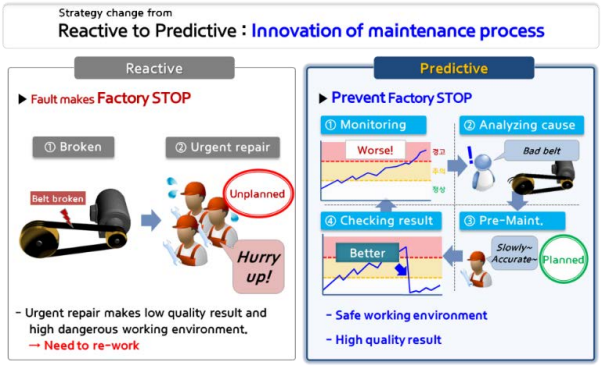


Figure 1. Strategy changes from reactive to predictive

Various sensing techniques such as temperature, pressure, flow, ultrasonic waves, vibration, and acoustic emission can be used to monitor the equipment condition. Among them, vibration monitoring can cover most of mechanical failures such as imbalance, mismatch, bearing defects, gears, looseness, noise, cracks, resonance, etc. [ISO 18436-2:2014] [Figure 2]

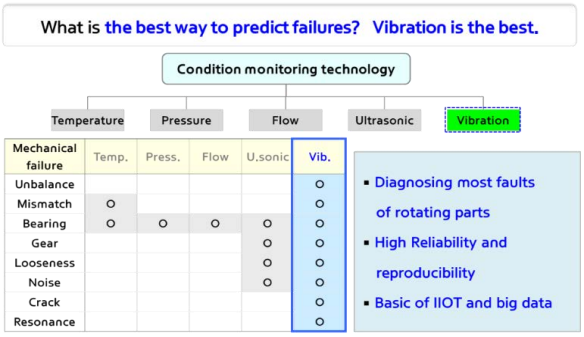


Figure 2. Comparison of Condition monitoring technology

For predictive maintenance, not all equipment is monitored. Instead, sensors, signal processing modules, and servers capable of collecting equipment status data are installed selectively, focusing on those that can lead to extended downtime failures. The primary target equipment includes conveyors, drop-lifters, air conditioner fans, robots, controllers, and communication systems. VMS, or Vibration Monitoring System, monitors vibration data at the equipment's drive in real-time and can predict failures in advance. This system comprises a sensor unit for real-time vibration measurement, a signal processing module, a server for data collection and storage, and a monitoring device for status display. It is implemented on main equipment causing extended downtime failures (downtime exceeding 60 minutes). Over 5,000 sensors are installed in more than 2,000 machines across Hyundai/Kia factories worldwide. All system servers are connected to the company network, allowing authorized executives and employees to access data conveniently from anywhere, enabling real-time monitoring of equipment installed worldwide [Figure 3].

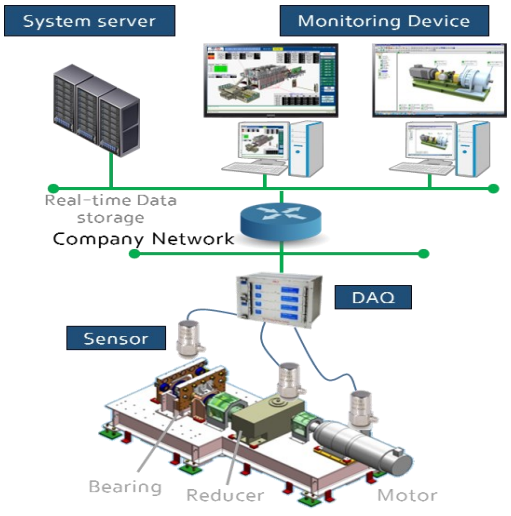


Figure 3. Vibration Monitoring System configuration

Similar to vibration monitoring, an approach to determine the number of shorted turns based on the difference between the phase current of the healthy and faulty machine is proposed [Nandi, S., Toliyat]. Condition monitoring of rotating machinery is crucial for ensuring operational reliability and preventing catastrophic failures. While vibration monitoring has long been the primary method for detecting mechanical faults, current monitoring offers a complementary approach with distinct advantages. Current monitoring is comparatively inexpensive to implement and is well-suited for diagnosing electrical faults in motors and inverters. This system also includes development and application of current-based condition monitoring technology to prevent electrical faults in rotating machinery. Technology for monitoring the current of the equipment's driving unit has been developed and implemented. However, challenges such as slow sampling of current data, the requirement for a large-capacity server, and a lack of diagnostic technology persist. To address this issue, a diagnostic function block capable of current analysis has been developed and deployed. Integrated within the PLC controller's code programs, this diagnostic function block receives various parameters including current, frequency, and control setting values of the equipment's driving unit as input. The output comprises diagnostic results, which are transmitted to the cloud server for analysis. The figure depicting the diagnostic function block is shown below [Figure 4].

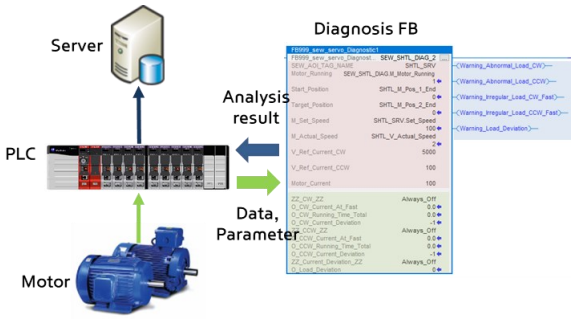


Figure 4. diagnostic function block for current monitoring

The increasing complexity and autonomy of robotic systems necessitate advanced condition monitoring techniques to ensure their reliable operation and prevent costly downtime. Traditional condition monitoring approaches often rely on centralized cloud-based data processing, which can lead to latency issues and communication bottlenecks. Edge computing offers a promising alternative by enabling real-time data analysis and decision-making at the edge of the network, closer to the data source [Niklas Tritschler].

This system involves installing vibration sensors on the motor and reducer of each axis of the robot and applying an autoencoder algorithm to assess anomaly scores. An edge

device was developed and utilized as a CMS (DAQ) to gather vibration data from the robot. Figure 5 illustrates the structure of robot vibration diagnosis system based edgeCMS system.

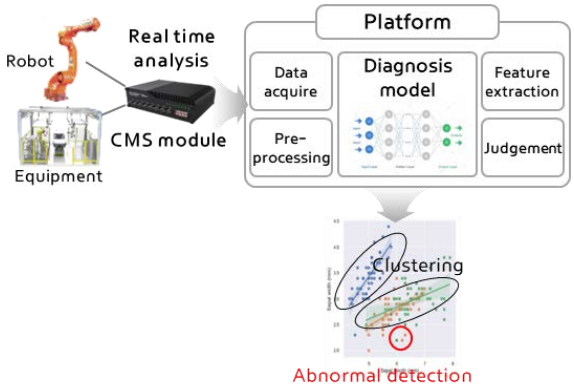


Figure 5. Robot vibration diagnosis system with edgeCMS

To monitor the condition of industrial robots, each manufacturer develops and provides a robot monitoring system. Hyundai(HRMS-Hyundai Robot Monitoring System), Yaskawa(Y-FAI) , Kawasaki(KRDS-Kawasaki Robot Diagnostic System), and Fanuc(ZDT-Zero Down Time system) robots are representative examples. The manufacturer's monitoring system typically displays basic operation information such as model and operation status graphs, along with notification history. It also provides alarm information and component replacement time when the reference value deviates from statistical norms. However, this system lacks failure prediction functionality. To address this, a Robot Predictive Maintenance System (RPMS) was developed. This system utilizes autoencoder models to learn normal states from manufacturer-provided robot monitoring data and identifies deviations from normal states [Figure 6].

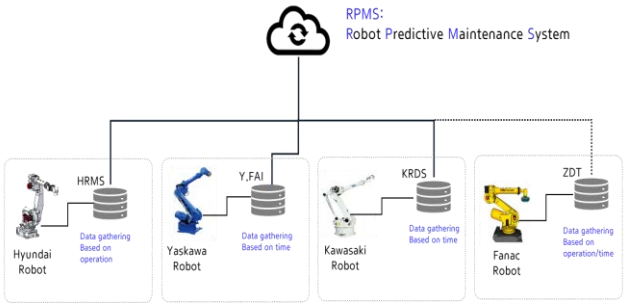


Figure 6. Robot monitoring system

The PLC(Programmable Logic Controller) program backup system is being implemented. It mainly consists of an agent PC managing program change points and a database storing the results. Hyundai/Kia factories employ PLCs from various manufacturers including Siemens, Rockwell, Mitsubishi, Fuji, and LS, and all systems managing program change points of these PLCs are in use. However, to address issues related to ineffective program backup when the agent PC is improperly managed, a cloud-based technology integrating agents and databases has been developed. This technology manages fluctuation points effectively. Figure 7 illustrates the structure of the PLC program backup system based on the cloud.

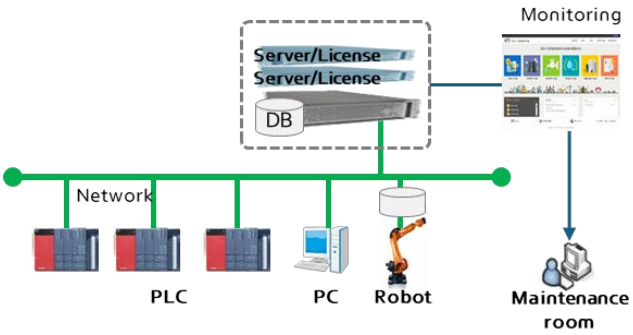


Figure 7. PLC program backup system based on cloud

2. INTEGRATED PHM SYSTEM

As the equipment monitoring system is currently implemented as separate systems, managing equipment becomes cumbersome. This involves checking multiple individual systems separately to diagnose a single piece of equipment, and there is no comprehensive system for analyzing equipment data. To address these challenges, a next-generation integrated Prognostics and Health Management (PHM) system is being developed. This integrated PHM system, based on existing accumulated monitoring technology and diagnostic expertise, is being deployed in new factories within the Hyundai/Kia company.

2.1. Configuration of Integrated PHM system

The aim of the integrated PHM system is to consolidate equipment failure data monitoring and enhance failure diagnosis and analysis capabilities. The system comprises sections for equipment monitoring, diagnostic report management, fault diagnosis algorithms, and data transmission/reception interfaces.

Equipment monitoring encompasses vibration, drive current, electrical equipment status, PLC program backup status, robot operation status, robot vibration status monitoring, alarm lists, and maintenance history inquiries. Additionally, it enables monitoring of equipment status trends and system resources through trend graphs.

Diagnostic report management is linked with equipment abnormality alarm event management, facilitates automatic report generation and email dispatch of diagnostic reports, and integrates with the prevention task instruction module in ERP(SAP).

The analysis algorithms promote advanced analysis techniques, including time series trend analysis, frequency/pattern analysis, harmonic/rotational speed analysis, automatic vibration state analysis, and automatic failure type determination algorithms that detect sudden state changes.

Utilizing the IoT(Internet of Things) platform and Hyundai/Kia's standard collection module, the system collects equipment state data and offers basic functions for efficient and reliable data management. These include alarm management, equipment registration, device control, data collection, and algorithm management.

Currently, the integrated PHM system monitors equipment/robot vibration status, status of current, control panel status, network switch status, robot status, and PLC backup status. However, it can accommodate new monitoring solutions for equipment condition diagnosis. Figure 8 illustrates the configuration of the integrated PHM system.

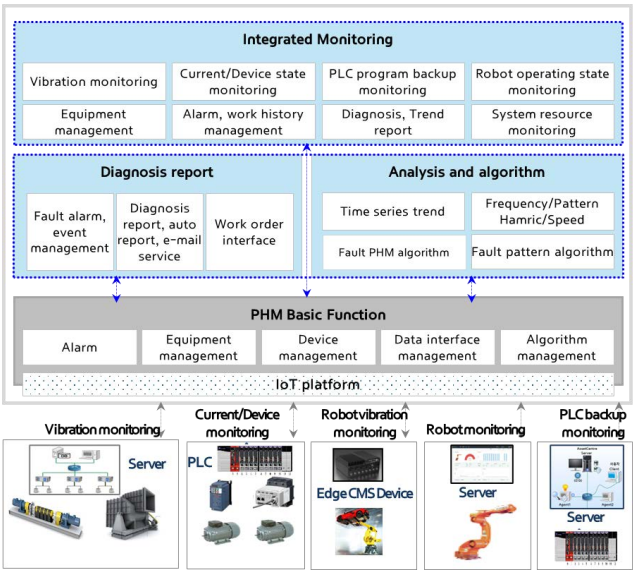


Figure 8. Configuration of integrated PHM system

Several data transactions and signal processing tasks are necessary for the integrated PHM system, including real-time data collection and analysis, CMS(Condition Monitoring System) device control, AI(Artificial Intelligence) algorithm execution, and accounting for network load and security considerations. Consequently, physical servers need to be configured for each factory.

The integrated PHM server comprises a web/app server, a DB(Database) server, and an Factory Talk Linx Gateway server(Rockwell) for interfacing with current and electronic component data. Figure 9 illustrates the configuration of the integrated PHM server.

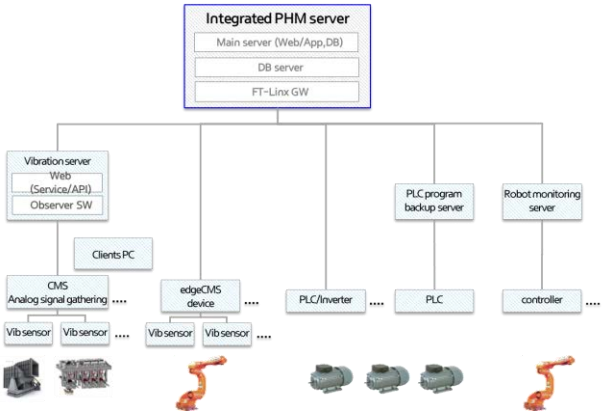


Figure 9. Configuration of integrated PHM server

2.2. Architecture of Integrated PHM system

The architecture of the integrated PHM system is designed to collect, analyze, and present equipment source data. Vibration monitoring, robot monitoring, and PLC backup utilize dedicated servers, with data stored on these servers interfacing with the integrated PHM server via a DB-to-DB interface method.

For collecting current/control panel state monitoring data, PLCs in the control panel gather and analyze necessary data using function blocks. This data is then collected by a dedicated collection server such as FTLinx GW and stored in the integrated PHM's InfluxDB using the OPC-UA protocol. FTLinx GW serves as a collecting tool for Rockwell systems, whereas different servers are required for PLCs from other manufacturers such as Siemens, LS, and Mitsubishi.

Robot vibration data is collected from the edge CMS device, and the analysis result is transmitted to the integrated PHM server through a file collection batch process. As the results from edgeCMS are stored as files, the system must possess the capability to collect and manage files.

The collected equipment status data undergoes backend analysis, including vibration and current abnormality diagnosis, diagnosis notification/report management, equipment status management, external system connection, data collection management, device/alarm management, and visualization data management.

At the front end, functions such as factory map-based equipment management, equipment status management/alarm management, and equipment-based information data visualization are implemented. Further details are provided in Figure 10 below.

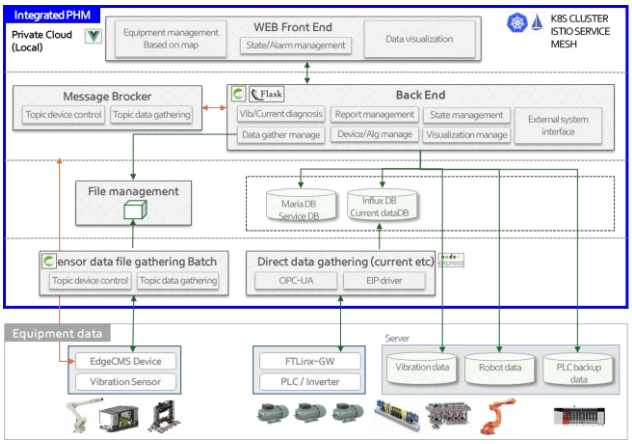


Figure 10. Architecture of Integrated PHM system

2.3. Monitoring contents and user interface

2.3.1. Main integral monitoring

The integrated system encompasses various monitoring purposes and incorporates a user-friendly UI/UX application, incorporating different factors based on a map interface. This enhances efficiency for maintenance engineers. Additionally, AI PHM algorithms are integrated into the system, enabling automatic vibration detection and spectrum analysis, serving as powerful tools. Automatic email notifications of abnormalities further enhance maintenance efficiency.

The main page of the integrated PHM system shows the equipment list, equipment status statistics (normal, caution, warning), daily alarm trends, shop-specific diagnostic result statistics, itemized diagnostic result statistics, and equipment status displayed on a map. Moreover, the system automatically presents analysis results on the main page to improve user intuition. Alarm history, diagnosis reports, action details, and maintenance work management content are also accessible on the main page. Specific and detailed results can be viewed by clicking on each menu. Figure 11 provides an example of the integrated main page.

Figure 11. Main page of Integrated PHM system

2.3.2. Vibration monitoring

The vibration monitoring analysis page configuration presents the average vibration level for each sensor along with the automatically calculated vibration variation. Sensors exhibiting significant changes are highlighted in red. Additionally, the page displays the equipment's vibration trend over time and automatically analyzed results of frequency spectrum analysis. For detailed spectrum analysis, 3D plots and heat maps are provided. Furthermore, the system features automatic generation and emailing of diagnostic reports [Figure 12].

Figure 12. Vibration monitoring page

2.3.3. Current and control panel state monitoring

For equipment current state monitoring, the system displays comprehensive analysis results of parameters such as current data and frequency data from the inverter driving the motor. These values indicate diagnostic results such as load current performance and fluctuations during machine operation. Additionally, the system enables monitoring of elements crucial for the inverter's lifespan, including IGBT, capacitor, and temperature. Sensor data for each location of the drop-lifter equipment, responsible for moving the car up and down, is also presented for monitoring sensor status. Furthermore, the system organizes a page to monitor power, temperature, and lifespan of the main elements in the control panel [Figure 13].

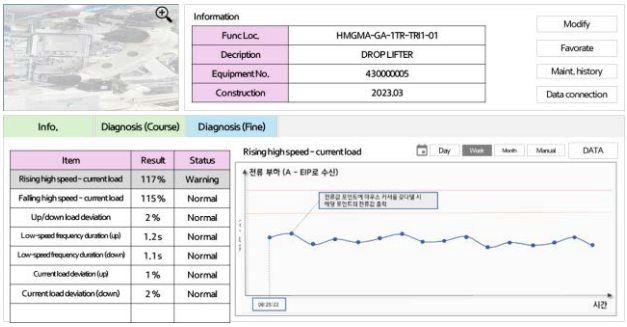


Figure 13. Current/Control panel monitoring page

2.3.4. Robot vibration monitoring

The method for diagnosing the vibration state of robots differs from that of general equipment described earlier. Unlike conventional equipment, robots lack distinct movement patterns and have short constant speed sections, making diagnosis challenging with traditional vibration analysis methods. To address this, a vibration sensor is attached to each axis of the robot, and an AI algorithm, specifically the auto-encoder method, is employed to predict signal outliers. By utilizing an auto-encoder, the difference between normal and abnormal data can be transformed into a health index score, facilitating equipment state trend prediction. The figure below illustrates the monitoring of the robot's vibration status [Figure 14].

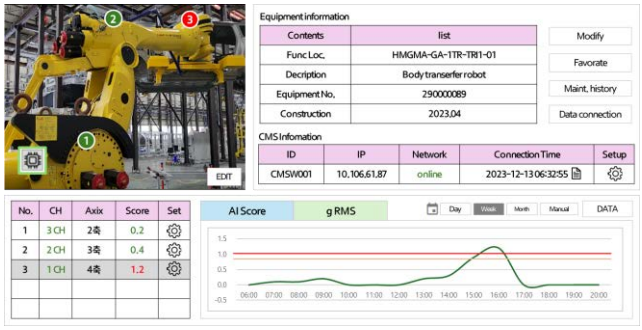


Figure 14. Robot vibration monitoring page

2.3.5. Robot operation monitoring

In the automotive manufacturing factory, there are 600 industrial robots in the vehicle welding factory, 200 in the painting factory, and 100 in the assembly factory. A robot monitoring system is developed and installed in each factory to monitor the operation information and condition of the robots. Since monitoring the operation status, alarm history, and error information of hundreds of robots in individual systems is challenging, the system is configured to initially display key results in coordination with the integrated PHM system, allowing users to review detailed information in individual robot monitoring systems as needed [Figure 15].

Figure 15. Robot operation monitoring page

2.3.6. PLC program backup monitoring

The equipment controller utilizes PLC, primarily employing ladder programs. When equipment operation is altered, monitoring the normality of program backups is managed through the PLC backup status inquiry page. This page displays the location, equipment name, and final backup date, and indicates any communication issues or conditions if backups are not performed. The image below depicts the PLC backup status inquiry page [Figure 16].

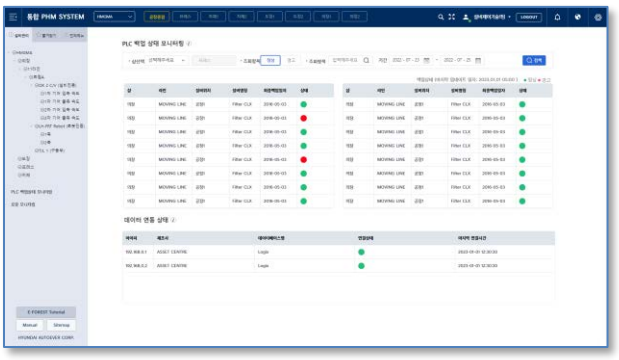


Figure 16. PLC backup monitoring page

3. Practical Application

Various PHM technologies for monitoring and predicting equipment conditions are being implemented in production plants. Initially, an integrated PHM system, integrating these technologies, is being deployed in new factories within Hyundai/Kia. These include the HMGMA Plant under construction in Savannah, Georgia, as well as the Gwangmyeong EVO Plant, Ulsan EV Plant, and Hwaseong EVO Plant in Korea. Subsequently, monitoring will extend to enhance equipment uptime across mass production plants. Equipment monitoring and technical support are facilitated by the Equipment Monitoring Center at Hyundai Motor's Ulsan plant [Figure 17].

Figure 17. Equipment Monitoring Center in Hyundai motor.

4. Conclusion

Vibration and current monitoring are underway for 2,000 rotating equipment units throughout Hyundai/Kia factories.

An integrated PHM system is being developed to streamline equipment management and analysis. It integrates data from various monitoring systems, including vibration, PLC, and robot diagnostics, facilitating trend detection and fault diagnosis. The system architecture encompasses components for data management, diagnostic reporting, and external system integration. We plan to continue activities using the integrated PHM system to efficiently monitor equipment status and dramatically improve downtime.

REFERENCES

Paulina Gackowiec (2019). General overview of maintenance strategies – concepts and approaches. Multidisciplinary Aspects of Production Engineering 2(1):126-139 DOI:10.2478/mape-2019-0013

Sudhanshu Goel (2022). A Methodical Review of Condition Monitoring Techniques for Electrical Equipment. papers.phmsociety.org

ISO 18436-2:2014 Condition monitoring and diagnostics of machines — Requirements for qualification and assessment of personnel — Part 2: Vibration condition monitoring and diagnostics

Nandi, S., Toliyat, H.A. and Li, X. (2005) Condition Monitoring and Fault Diagnosis of Electrical Motors—A Review. IEEE Transactions on Energy Conversion, 20, 719-729

Niklas Tritschler, Andrew Dugenske, Thomas Kurfess. (2021). An Automated Edge Computing-Based Condition Health Monitoring System: With an Application on Rolling Element Bearings. Journal of Manufacturing Science and Engineering. Jul 2021, 143(7): 071006 (8ps)

BIOGRAPHIES

Deog Hyeon Kim He is a technical senior manager with equipment control engineering team in Hyundai motors. His research interests include PHM of electronics and mechanics of automotive manufacturing industry and its artificial intelligent application. He leads PHM technology and operation in Hyundai/Kia company. He received the Master of Information and Communication Engineering, Gwangju Institute of Science and Technology, South Korea, in 2005. Since January 2014, he has been with Hyundai Motors. he acquired ISO 18436-1 category 3 in 2017.

Gun Sik Kim He works as e in the equipment control engineering team, his interested area is automatic spectrum analysis and installing integral monitoring IT system. He received the bachelor's degree in electrical and electronic Computers from Kyungpook National University, South Korea, in 2015. Since January 2015, he has been with Hyundai Motors. he acquired ISO 18436-1 category 3 in 2017.

Jung Ho Nam He works as an expert who analyze current and control panel data. And he is developing new control technology replacing PLC system in factory. He received the Bachelor of Electronic Engineering, Aju University, South Korea, in 2006. Since January 2006, he has been with Hyundai Motor. He has been with Hyundai Motors. he acquired ISO 18436-1 category 2 in 2017.

Jin Woo Park He works as team leader of equipment control engineering team. He received the Bachelor of Electronic Engineering, Dong-A University, South Korea, in 2002. Since January 2003, he has been with Hyundai Motor.

Active learning for gear defect detection in gearboxes

Wenzhi Liao[1,2], Roeland De Geest[1], Djordy Van Maele[3], Jean Carlos Poletto[3], Laveen Prabhu Selvaraj[4], Ted Ooijevaar[1], Luk Geens[4]

[1] *Flanders Make, Oude Diestersebaan 133, 3920 Lommel, Belgium*
Wenzhi.liao@FlandersMake.be

[2] *IPI-TELIN, Ghent University, St-Pietersnieuwstraat 41, B-9000 Gent, Belgium*

[3] *Ghent University, Soete Laboratory, Technologiepark Zwijnaarde 46, 9052 Zwijnaarde, Belgium*

[4] *ZF Wind Power Antwerpen NV, Gerard Mercatorstraat 40, 3920 Lommel*

ABSTRACT

Condition monitoring of gears in gearboxes is crucial to ensure performance and minimizing downtime in many industrial applications including wind turbines and automotive. Monitoring techniques using indirect measurements (i.e. accelerometers, microphones, acoustic emission sensors and encoders, etc.) face challenges, including the defect interpretation and characterization. Vision-based gear condition monitoring, as a direct method to observe gear defects, has the capability to give a precise indication of the starting point of a potential surface failure, but suffers from the image annotations (to train a reliable vision model for automatic defect detection of gears). In this paper, we propose an active learning framework for vision-based condition monitoring, to reduce the human annotation effort by only labelling the most informative examples. In particular, we first train a deep learning model on limited training dataset (annotated randomly) to detect pitting defects. To select which samples have the highest priority to be annotated, we compute the model's uncertainty on all remaining unlabeled examples. Bayesian active learning by disagreement is exploited to estimate the uncertainty of the unlabeled samples. We select the samples with the highest values of uncertainty to be annotated first. Experimental results from defect detection of gears in gearboxes show that with less than 6 times image annotations, we can achieve similar performances.

1. INTRODUCTION

Detecting defects on gear surfaces is essential for maintaining the safety, performance, and longevity of machinery, while also ensuring quality control and minimizing downtime and costs, especially for gearboxes in high-power-density machines (e.g., wind turbines). Many approaches exploit indirect measurements acquired from accelerometers, microphones, acoustic emission sensors and encoders to monitor the damage evolution in gears (Surucu, Gadsden, & Yawney, 2023; Feng, Ji, Ni, & Beer, 2023). However, this indirect way of gear condition monitoring (e.g., vibration analysis) suffers from relative indicators and setting good thresholds to accurately track the gear damage (Surucu et al., 2023). Moreover, the indirect measurements cannot well characterize the defects (e.g., size, location, type) of the gears (Van Maele et al., 2023). Vision monitoring, which is a direct method to observe defects has the capability to give a precise indication of the starting point of a potential surface failure. Gear damage is often validated using visual inspection with borescopes or fibre scopes. However, such a system is used in some domains (mainly in wind turbines) as a periodic maintenance procedure but expensive equipment and permanent machine stop is needed (Coronado & Fischer, 2015). Recent advances in computer vision and machine learning have revolutionized industrial maintenance practices, allowing for the development of automated systems capable of visually inspecting and analyzing gear surfaces. Vision-based approaches utilize cameras and sensors to capture images or videos of gears during operation, enabling the extraction of meaningful visual features for condition assessment (Allam, Moussa, Tarry, & Veres, 2021; Qin, Xi, & Chen, 2023; Miltenović, Rakonjac, Oarcea, Perić, & Rangelov, 2022). This shift towards visual inspection not only facilitates continuous monitoring but also provides a more comprehensive understanding of gear health by capturing subtle surface details and anomalies. Massive image data can be acquired by high-speed cameras for visual condition monitoring of gears. Deep learning, particularly convolutional neural networks (Allam et al.,

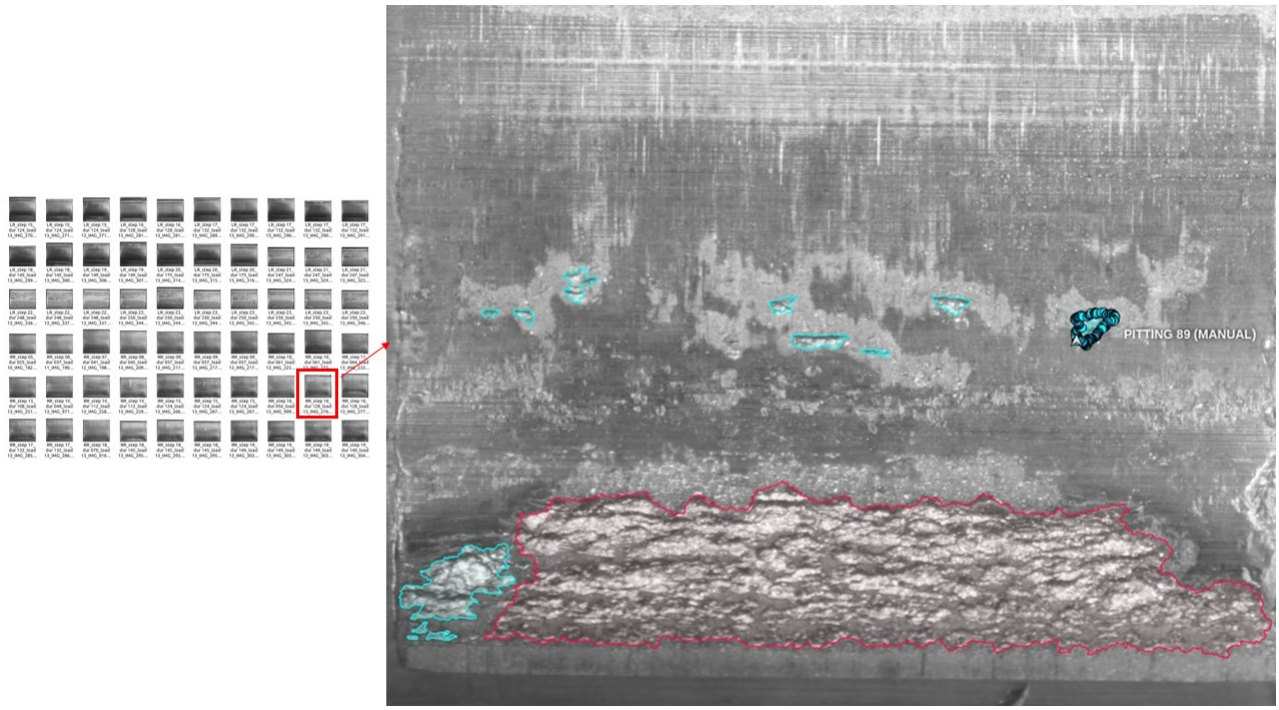


Figure 1. More than 20 minutes were taken by our experts to annotate defect of pitting in a single frame of image. Two types of defects were annotated for all images: micropitting (red), and pitting (cyan).

2021), has shown remarkable success in image-based tasks, making it well-suited for gear surface defect detection. However, to train a reliable vision deep learning model for automatic defect detection of gearboxes, a huge amount of image data typically needs to be annotated, which is expensive and time-consuming (Alzubaidi, Bai, Al-Sabaawi, & et al., 2023). For example, it takes more than 20 minutes to annotate all pitting defects in a single frame of image, as shown in Figure 1. Moreover, image datasets acquired during full lifetime degradation tests datasets contain many similar examples that bring no additional information to the diagnostic model. To overcome these problems, the active learning method was exploited to select the most informative indirect signals (e.g., vibration, supervisory control and data acquisition) for gearbox fault diagnosis (Chen et al., 2019) or wind turbine condition monitoring (Bao, Zhang, Hu, Feng, & Liu, 2023). Recent work on vision-based defect segmentation also showed that active learning framework can reduce data storage and labeling costs for imbalanced industrial datasets (Li et al., 2023).

To reduce the cost on manual annotation, this paper proposes an active learning framework to address the challenge of acquiring labeled data by iteratively selecting the most informative images for annotation. To the best of our knowledge, this paper is the first study to apply deep active learning for vision-based gear defect segmentation/detection in gearboxes. Specifically, a few images (i.e. around 20) were initially annotated to train a deep learning model for defect detection. To choose which gear images will be the first priority to be annotated, we then compute the model's uncertainty on all remaining unlabeled examples, where Bayesian active learning (Atighehchian et al., 2022) by disagreement is exploited to estimate the uncertainty of the unlabeled samples. The samples with the highest values of uncertainty will be chosen to be annotated first. We repeat the image annotations iteratively (e.g., top 10 images ranking according to the uncertainty will be annotated in each iteration) until we achieve a satisfactory performance.

The structure of this paper is as follows. Section 2 introduces the active learning framework. Section 3 details the experimental data collection and processing. The experimental results of defect detection on gear flanks are presented and discussed in Section 4. Finally, the conclusions of this paper are drawn in Section 5.

2. METHODOLOGY

2.1. Deep segmentation model

To monitor the damage evolution in gears, our solution first segments the damaged regions (defect) in the acquired images, then characterizes these damaged regions (change of size, shape, depth, etc.). A Python library with Neural Networks for Image Segmentation based on PyTorch (SMP) (Iakubovskii, 2019) is exploited for defect segmentation task

in this paper, as it is an open-source library built on top of PyTorch, specifically tailored for semantic segmentation tasks in computer vision. Semantic segmentation involves assigning a class label to each pixel in an image, thus dividing the image into distinct regions corresponding to different object classes. Semantic segmentation is additionally assigning each detected object a category and discriminates between objects of the same category. SMP includes an efficient and flexible implementation of Feature Pyramid Network (FPN) (Lin et al., 2017) for semantic segmentation tasks, combining low- and high-resolution features via a top-down pathway to enrich semantic features at all levels (multi-scale features). By leveraging multi-scale features and transfer learning, SMP-FPN enables accurate and robust segmentation of objects in images across various scales and contexts, fitting perfectly with the defect detection in the gears (defect area sizes changing).

The initial training dataset is very limited, since image annotation of these defect in the gears are challenging and time consuming. Therefore, we leverage pre-trained weights from models trained on large-scale image datasets such as ImageNet. The pre-trained weight of ResNet-18 (He, Zhang, Ren, & Sun, 2016), with a convolutional neural network that is 18 layers deep[1], is exploited in our segmentation model. The pre-trained model has been previously trained on more than a million images from the ImageNet database and contains the weights and biases that represent the features of whichever dataset it was trained on. These low-level learned features are often transferable to different data, including gears. For example, a model trained on a large dataset of natural objects (e.g., bird, fish images) will contain learned features like edges or textures that would be transferable defects in gears, which helps improve the performance of the segmentation model (especially with very small training sample size).

2.2. Active learning for image annotation

Even with a pre-trained model, the segmentation performances are still poor, especially for images mixed with two classes of "micropitting" and "pitting", as shown Figure 2, regions of micropitting were misclassified into pitting (poor performances in confusion matrix), while pitting defects were misclassified into background. An easy and simple solution to improve the performances is to add more annotated images into the training dataset. With a high-speed camera, we can acquire more than 60 image per second, around 30,000 images for 8 hours. However, image annotation is time consuming for our experts (an image shown in Figure 1 may take 20 minutes to annotate), even with advanced annotation tool CVAT [2]. Since it is infeasible for an expert to annotate all the acquired images, two challenges need to be solved: (1) which images should be first annotated? (2) how many images should be annotated for a reliable prediction?

Active learning aims to minimize the annotation effort required by selecting the most informative samples for annotation, i.e., the samples that would most increase the model accuracy. Active learning is a machine learning paradigm where a model iteratively queries the user or a human annotator for the labels of the most informative samples. This can lead to significant savings in time and resources compared to traditional approaches that rely on labeling large amounts of data upfront or passive learning from a fixed dataset.

Many active learning approaches have been proposed (Beluch, Genewein, Nürnberger, & Köhler, 2018; Kirsch, Amersfoort, & Gal, 2019; Wan et al., 2023), but some of these methods are either not scalable to large datasets or too slow to be used in a more realistic environment (e.g., in a production setup) (Atighehchian, Branchaud-Charron, & Lacoste, 2020). We exploit Bayesian Active Learning by Disagreement (BALD) (Atighehchian et al., 2020) in this paper to select the most informative samples for annotation. BALD leverages Bayesian modeling to estimate the uncertainty of a predictive model and selects samples where the model's predictions are most uncertain. In particular, BALD involves calculating the mutual information between the model's predictions and the model's parameters, given the observed data. Let D denote the labeled dataset, where $D = (\boldsymbol{x}_i, y_i)_{i=1}^{N}$ with inputs $\boldsymbol{x}_i$ and corresponding labels y_i. Let θ represent the model parameters, and $f_\theta(\boldsymbol{x})$ denote the predictive distribution of the model. The BALD acquisition function is defined as the mutual information between model parameters and potential labels of unlabeled data $\boldsymbol{x}$:

$$\begin{aligned} BALD(x) &= \boldsymbol{I}[y, f_\theta(\boldsymbol{x})] \\ &= \boldsymbol{H}[y] - \boldsymbol{E}_{p(f_\theta(\boldsymbol{x})|D)}[\boldsymbol{H}[y|f_\theta(\boldsymbol{x})]] \end{aligned} \tag{1}$$

Where:

- $\boldsymbol{I}[y, f_\theta(\boldsymbol{x})]$ is the mutual information between the label y and the model's prediction $f_\theta(\boldsymbol{x})$ for an unlabeled data point $\boldsymbol{x}$.
- $\boldsymbol{H}(y)$ is the entropy of the label distribution, measuring uncertainty in the label predictions.
- $\boldsymbol{E}_{p(f_\theta(\boldsymbol{x})|D)}$ is the expectation over the posterior distribution of the model given the current dataset D
- $\boldsymbol{H}[y, f_\theta(\boldsymbol{x})]$ is the conditional entropy of the label distribution given the model's prediction.

Intuitively, samples with higher BALD scores are those for which the model's predictions are most uncertain and thus are most informative for learning. By querying such uncertain samples, the model can learn more effectively with limited annotated data, leading to efficient data annotation for model training. In a normal image annotation task, our expert

[1] https://www.kaggle.com/datasets/pytorch/resnet18
[2] https://github.com/opencv/cvat

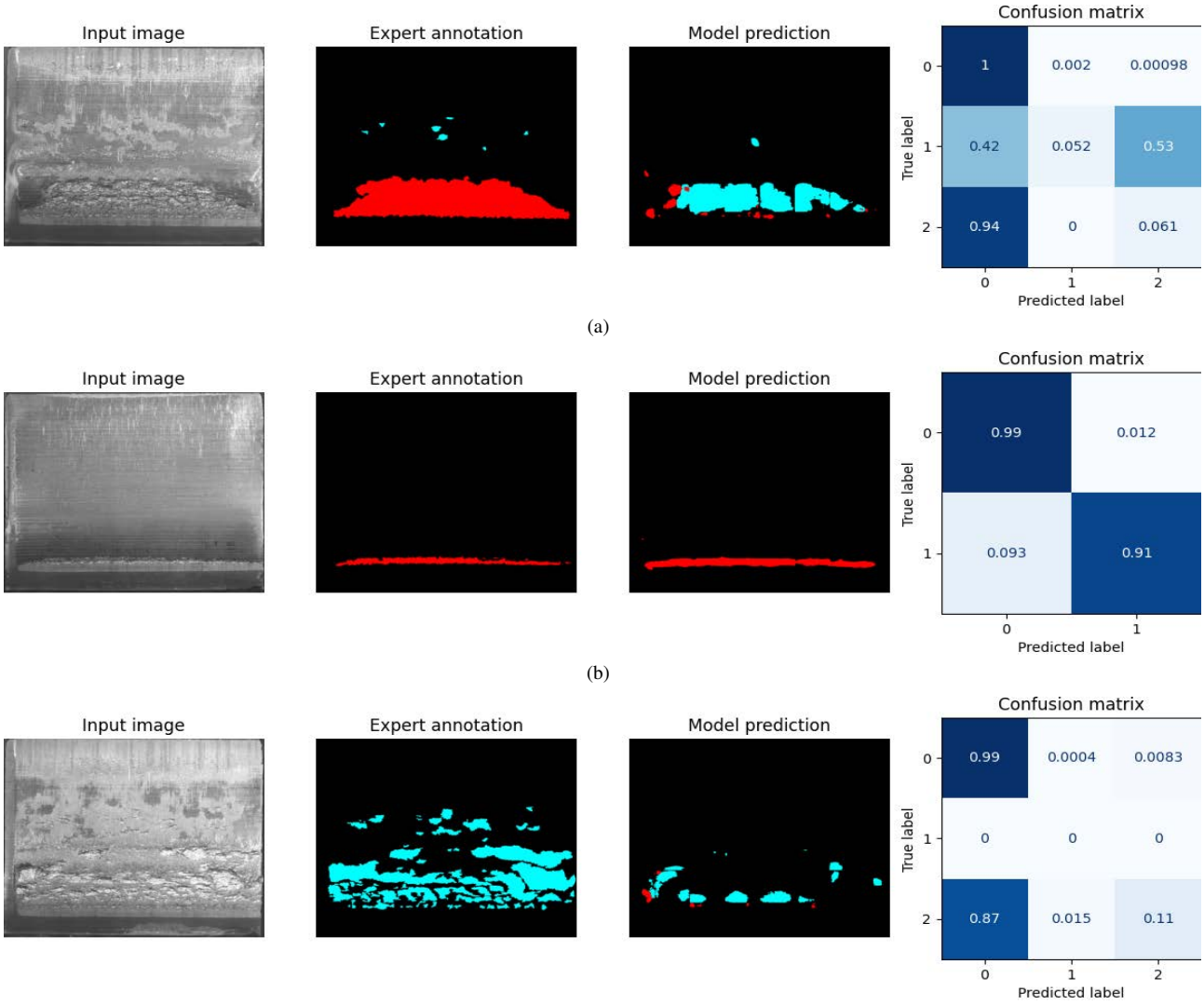


Figure 2. Performances on segmentation model trained on a small training dataset. The confusion matrices in the right column show the performances for three test images, 3 classes were defined by our experts in the images, with class label (color) 0: background (dark), 1: micropitting (red), 2: pitting (cyan).

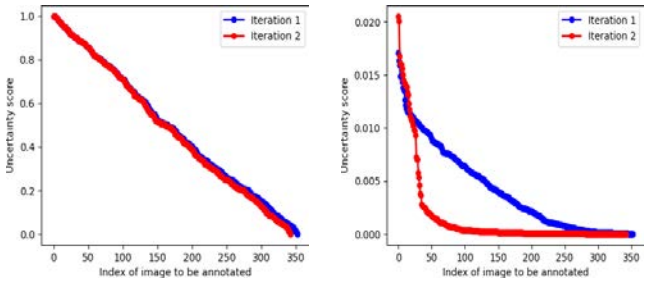
annotators will start annotate images according to their order uploaded into a annotation tools (CVAT) or the project coordinator will assign a certain number of images randomly to each annotator. Compared to the active learning with random selection of samples for annotation, the uncertainty score of active learning with BALD tends to zero when reaching to 300 images in the first iteration, as shown in Figure 3. The active learning process using BALD is iterative. After annotating the selected samples and incorporating them into the training set, the model is retrained, and the process repeats. Over multiple iterations, the model becomes increasingly accurate, and the uncertainty decreases, leading to more confident predictions. The annotation loops will stop until the end-users satisfy with the performances, which can be evaluated by either through a matrix on validation dataset, or by manually interpretation on randomly selected images (if not enough validation reference images). Figure 3(b) shows that the uncertainty score of active learning with BALD tends to zero after 180 images in the second iteration, while active learning with random sample selection still needs to annotate all images to achieve this. By focusing on samples where the model's predictions are most uncertain, BALD enables efficient learning with limited annotated data.

Table 1. Acquired images and manual annotations.

No. Teeth	No. Annotated teeth	No. Images	No. Annotated Images	No. Annotated Polygons
54	18	1370	438	1036

Table 2. Data split (within 18 annotated teeth, 438 annotated images) for active learning.

Initial training dataset	Tooth 1 (23 images)
Validation dataset	Tooth 15 (31 images)
Test dataset	Tooth 5 (31 images)
Pool dataset	The other 15 teeth (353 images)



(a) (b)

Figure 3. Uncertainties scores by active learning with: (a) random selecting samples for annotation, (b) BALD. Iteration 1 means 10 annotated images are added into the initial training dataset and retrain the model of active learning; iteration 2 means 20 images are iteratively annotated and added into the initial training dataset.

3. Experiments Data Collection and Processing

3.1. Data Collection

A dataset containing images of gears subjected to accelerated lifetime tests was provided by ZF Wind Power. While a brief description of the dataset acquisition is present here, the reader is referred to the work of (Boemher, 2019) for further details. The dataset consists of two accelerated lifetime tests, and was generated on a standard FZG[3] gear test rig at ZF Wind Power. At selected moments of the test, the equipment was stopped and images of both meshing gears were manually captured using a Canon EOS 500D camera. Figure 4 illustrates the gear degradation of a gear flank throughout the test. Two pairs of standard FZG C14 spur gears with 16 teeth (pinion) and 22 teeth (wheel) were tested on each accelerated lifetime test. In the first test, with total duration of $152h$, the test was stopped 31 times for acquiring the image of the gear flanks. Meanwhile, on the second test with total duration of $250h$, image acquisition was performed 23 times. A prior qualitative assessment determined that the wheel of the first test did not developed damage. Hence, the assessed dataset is composed of 54 teeth: pinion (16) of first test, plus pinion (16) and wheel (22) of the second test.

[3]Forschungsstelle für Zahnräder und Getriebebau, which denotes the Gear Research Center at the Technical University of Munich

3.2. Experimental Setup

The accelerated lifetime testing procedure was designed to generate micropitting and pitting wear on the visually monitored gear surfaces. As shown in Table 1, 54 teeth were used in experiments and a large amount of images was acquired by our camera. After filtering the unclear images (i.e., blurry, noisy, etc.) and pre-processing, we obtain 1370 images, of which 438 images were annotated by our experts, as shown in Table 2. The annotation effort varied according to the amount of defects in each image, taking approximately 60 hours to fully annotate the dataset (438 images), and in some cases up to 30 minutes were required to annotate a single frame.

Two sets of experiments are compared:

- **Fixed_SMP**: train the SMP-FPN models using Initial training dataset + Pool dataset, totally 16 teeth, 376 annotated images;
- **Active_SMP**: train the SMP-FPN model initially on Initial training dataset (Tooth 1, with 23 annotated images in total)

Then a number of annotated samples (i.e. 10 in each iteration) selected by active learning from the Pool dataset are iteratively added into the training dataset, and the model is retrained. Within active learning segmentation, we will compare different methods to select samples for first priority to be annotated, such as:

- **Active_SMP_Random**: select 10 images randomly in each iteration;
- **Active_SMP_BALD**: select 10 images by using BALD method in each iteration.

We set some parameters for model training as: batch size: 8, epochs: 100, learning rate: 0.0001. To reduce inherent randomness in the training of deep networks, each experiment runs five times for active learning.

For performance evaluations, we exploit the confusion matrix (Powers, 2011) to report the performance of a segmentation model on a single image. This confusion matrix helps in understanding where the segmentation model is making errors, whether it is under-segmenting or over-segmenting certain classes, or if there are misclassifications between classes. It is an essential tool for evaluating the effectiveness of segmentation algorithms. To evaluate the segmentation models on the whole test dataset, we exploit mean intersection over

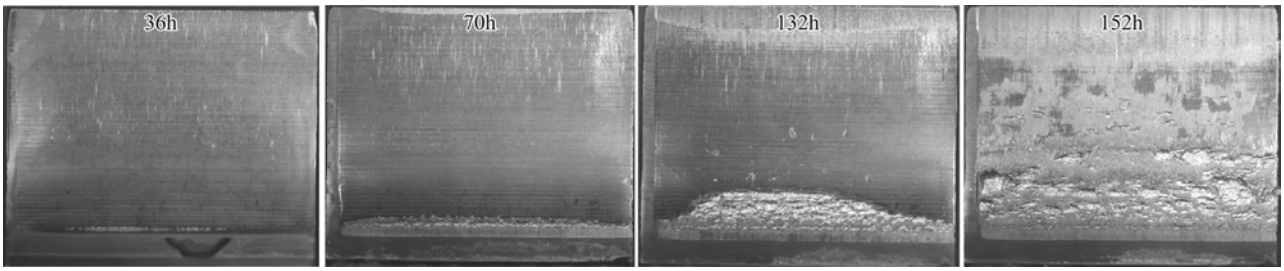


Figure 4. Example of images collected at selected moments of the test, showing the evolution on the gear degradation with the test duration.

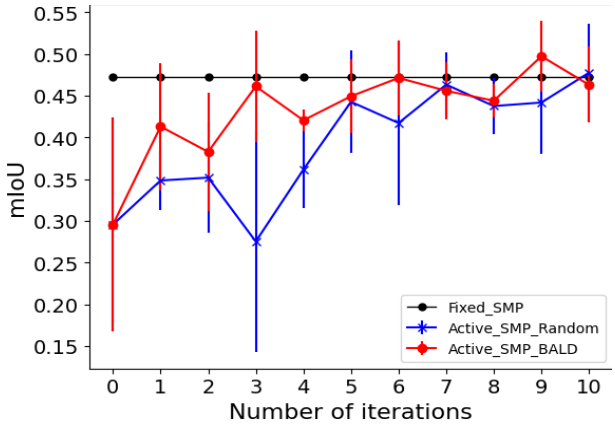


Figure 5. Mean and standard deviation mIoU of different segmentation methods. Note: the Fixed_SMP method is trained by using fixed number of 376 annotated images; iteration 0 means the model of active learning is trained on Initial training dataset (Tooth 1 with 23 images); 10 images will be selected in each iteration for annotation and then added into the training dataset, the model with active learning will be retrained (for example iteration 3 means 30 images are iteratively annotated and added into the Initial training dataset). We repeated the active learning segmentation experiments 5 times.

union (mIoU), which measures the overlap between the predicted segmentation and the ground truth segmentation for each class or object in the image (with Python implementation[4]). The value of the metric ranges from 0 to 1, higher value indicates better performance on segmentation.

4. Results and Discussions

Figure 5 compares the performances of segmentation models trained by Fixed_SMP and Active_SMP. The changes of the predicted segmentation maps by adding more annotated images into the training dataset can be found in Figure 6. We take several test images as examples and show the segmentation results and their confusion matrices by using Fixed_SMP and active SMP_BALD in Figure 7.

Based on Figure 5, we can find that with 53 annotated images (i.e. 3 iterations), active learning with BALD can achieve similar performances as Fixed_SMP (where more than 360 annotated images are used for training), which requires 6 times less annotated images for training, reducing more than 6 times the manual annotation effort. Moreover, Active_SMP_BALD performs better than Active_SMP_Random, especially for the first 5 iterations, when a small number of images are selected for annotations. This means that BALD can select the most informative images (out of a large dataset) for annotation when limited manpower is available for annotation. The model can learn more effectively with fewer BALD selected images, leading to efficient data annotation for model training. As more annotated images (more than 70 annotated images) are added into the training dataset, Active_SMP_Random converges to similar performances as the method of Fixed_SMP, indicating the redundancy in the image annotations. This is because images acquired during full lifetime degradation for multiple teeth contain many similar defects that bring no additional information for model training.

The segmentation models with active learning becomes more stable, as more annotated images added into the training dataset, as indicated by the changes of standard deviation in Figure 5. However, there are scenarios where increasing the training sample size (by adding more annotated images) might seemingly degrade segmentation performance, defects of “micropitting” appear in Figure 6-7 (Fixed_SMP for the third image) as more annotated images added. This may be due to overfitting, the model should generalize better with more training samples (that have similar distributions as the test images). One solution is to add more images that are representative of different classes into the pool dataset for active learning.

Compared to the human expert annotations, the segmentation results predicted by deep learning models (even for Fixed_SMP) need to be improved, regions of “micropitting” and “pitting” are misclassified into background, whereas some background regions are also misclassified into “micropitting”, as indicated by the confusion matrices in Figure 7. Challenges remain in predicting very tiny ”pitting” defects, as well as images mixed with big “micropitting” defects and

[4]https://lightning.ai/docs/torchmetrics/latest/segmentation/mean_iou.html

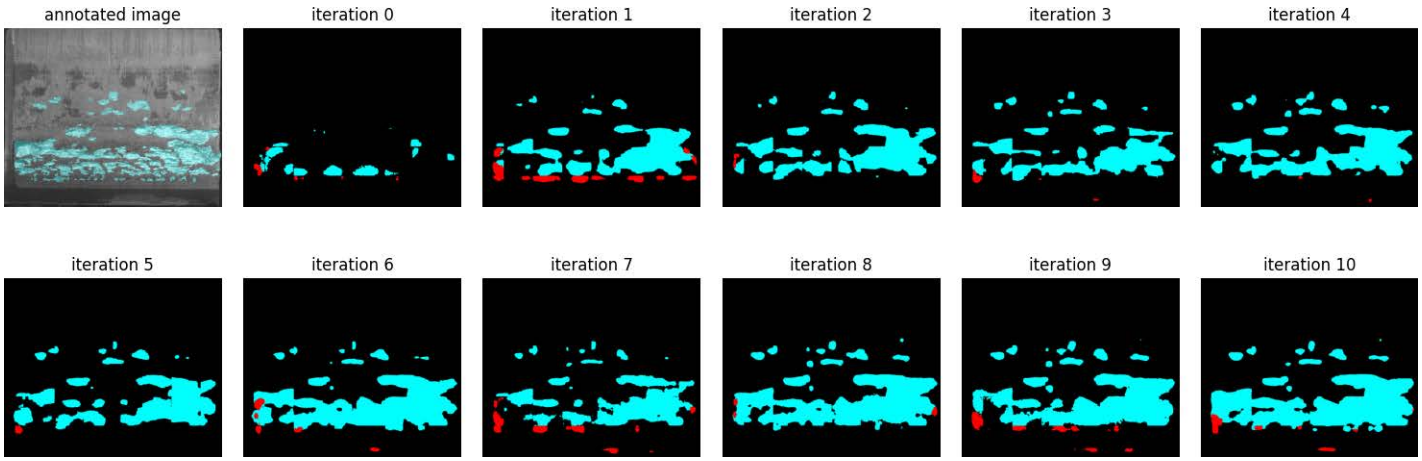
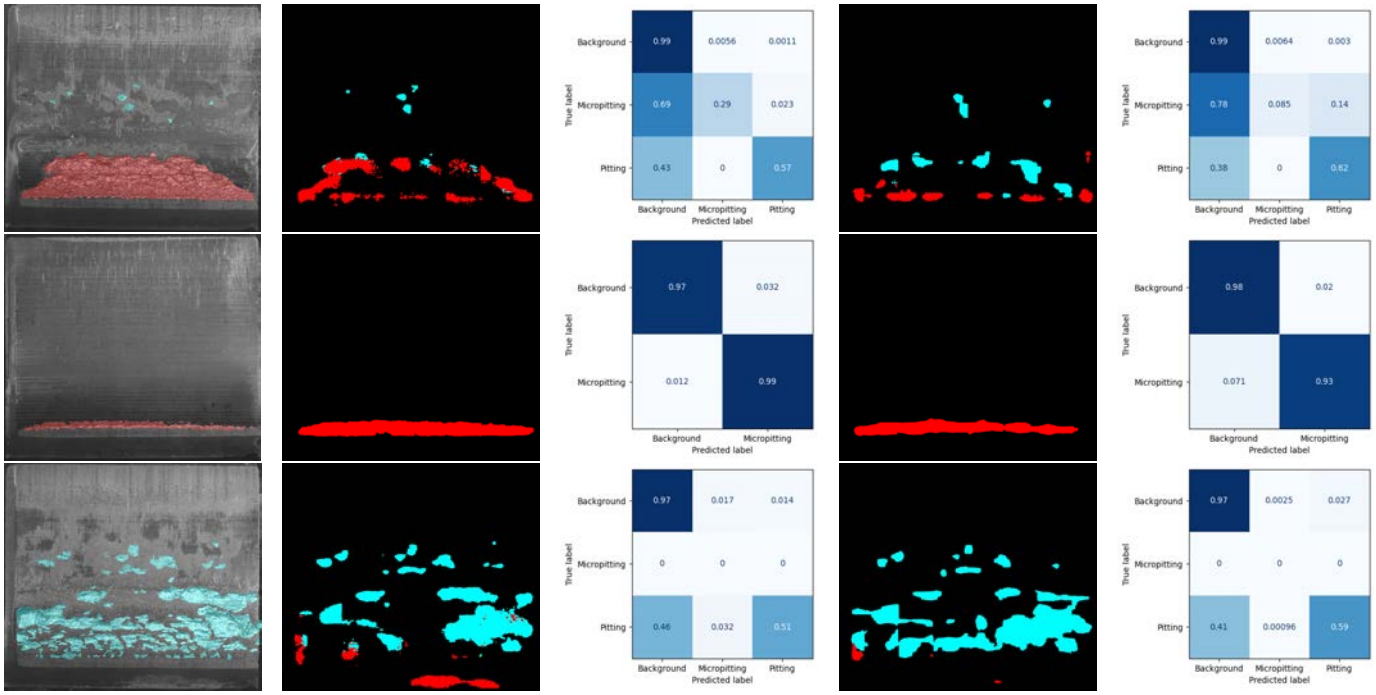


Figure 6. Prediction map changes as adding more annotated images (selected by BALD) into training dataset. 10 images will be annotated in each iteration and then added into the training dataset.



Expert annotations | Fixed_SMP | Fixed_SMP | active_SMP_BALD | active_SMP_BALD

Figure 7. Performances on segmentation by fixed training number VS. active learning. Each row has one test image, column 1 shows highlighted annotated images by experts, column 2 and 3 show predicted segmentation maps and confusion matrices by Fixed_SMP, column 4 and 5 are predicted segmentation maps and confusion matrices by Active_SMP_BALD with three iterations (53 training images).

tiny ”pitting” defects.

5. Conclusions

This paper focus on training a reliable deep learning segmentation model for defect detection in gears using less image annotations. In particular, Bayesian Active Learning by Disagreement (BALD) was exploited to select the most informative images for annotation iteratively until the satisfied performances were achieved. Experimental results show that with less than 6 times image annotations, we can achieve similar performances, leading to significant savings in time and resources compared to traditional approaches that rely on labeling large amounts of data upfront. However, gear surfaces exhibit a variety of defect types and patterns, and the successful identification of these defects requires a model capable of learning intricate features and subtle variations. The initial results in this paper can be extended by considering: (1) uncertainties from human annotations (annotations may be different by different human annotators in Figure 8), (2) imbalance in class distribution (some classes have more annotations than the other classes), (3) data augmentation to increase diversity in the training image dataset, (4) validation of the active learning methods on wider applications using some public datasets (e.g., ball screw drive surface defect dataset (Schlagenhauf & Landwehr, 2021)) for more comprehensive comparisons.

Acknowledgment

This research was supported by Flanders Make, the strategic research centre for the manufacturing industry, and more precisely the SBO (Strategic Basic Research) project for QED (Quantified Evolution of Degradation in gears, NO.: 2020-1138). The authors would like to thank ZF Wind Power NV for providing the gear image dataset that was used in this research.

References

Allam, A., Moussa, M., Tarry, C., & Veres, M. (2021). Detecting teeth defects on automotive gears using deep learning. *Sensors*, *21*(24).

Alzubaidi, L., Bai, J., Al-Sabaawi, A., & et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(46).

Atighehchian, P., Branchaud-Charron, F., Freyberg, J., Pardinas, R., Schell, L., & Pearse, G. (2022). *Baal, a bayesian active learning library.* `https://github.com/baal-org/baal/`.

Atighehchian, P., Branchaud-Charron, F., & Lacoste, A. (2020). *Bayesian active learning for production, a systematic study and a reusable library.*

Bao, C., Zhang, T., Hu, Z., Feng, W., & Liu, R. (2023). Wind turbine condition monitoring based on improved active learning strategy and knn algorithm. *IEEE Access*, *11*, 13545-13553.

Beluch, W., Genewein, T., Nürnberger, A., & Köhler, J. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 9368–9377).

Boemher, D. E. (2019, 4). Computer vision for gear alignment check and condition monitoring of wind turbine gearboxes.

Chen, J., Zhou, D., Guo, Z., Lin, J., Lyu, C., & Lu, . (2019). An active learning method based on uncertainty and complexity for gearbox fault diagnosis. *IEEE Access*, *7*, 9022-9031.

Coronado, D., & Fischer, K. (2015). Condition monitoring of wind turbines : State of the art , user experience and recommendations project report..

Feng, K., Ji, J. C., Ni, Q., & Beer, M. (2023). A review of vibration-based gear wear monitoring and prediction techniques. *Mech. Syst. Signal Process.*, *182*, 109605.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778).

Iakubovskii, P. (2019). *Segmentation models pytorch.* https://github.com/qubvel/segmentation_models.pytorch. GitHub.

Kirsch, A., Amersfoort, J., & Gal, Y. (2019). *Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning.*

Li, W., Li, B., Niu, S., Wang, Z., Liu, B., & Niu, T. (2023). Selecting informative data for defect segmentation from imbalanced datasets via active learning. *Advanced Engineering Informatics*, *56*, 101933.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (p. 936-944).

Miltenović, A., Rakonjac, I., Oarcea, A., Perić, M., & Rangelov, D. (2022). Detection and monitoring of pitting progression on gear tooth flank using deep learning. *Applied Sciences*, *12*(11).

Powers, D. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *ArXiv*, *abs/2010.16061*.

Qin, Y., Xi, D., & Chen, W. (2023). Gear pitting measurement by multi-scale splicing attention u-net. *Chinese Journal of Mechanical Engineering*, *36*(50).

Schlagenhauf, T., & Landwehr, M. (2021). Industrial machine tool component surface defect dataset. *Data in Brief*, *39*, 107643.

Surucu, O., Gadsden, S. A., & Yawney, J. (2023). Condition monitoring using machine learning: A review of the-

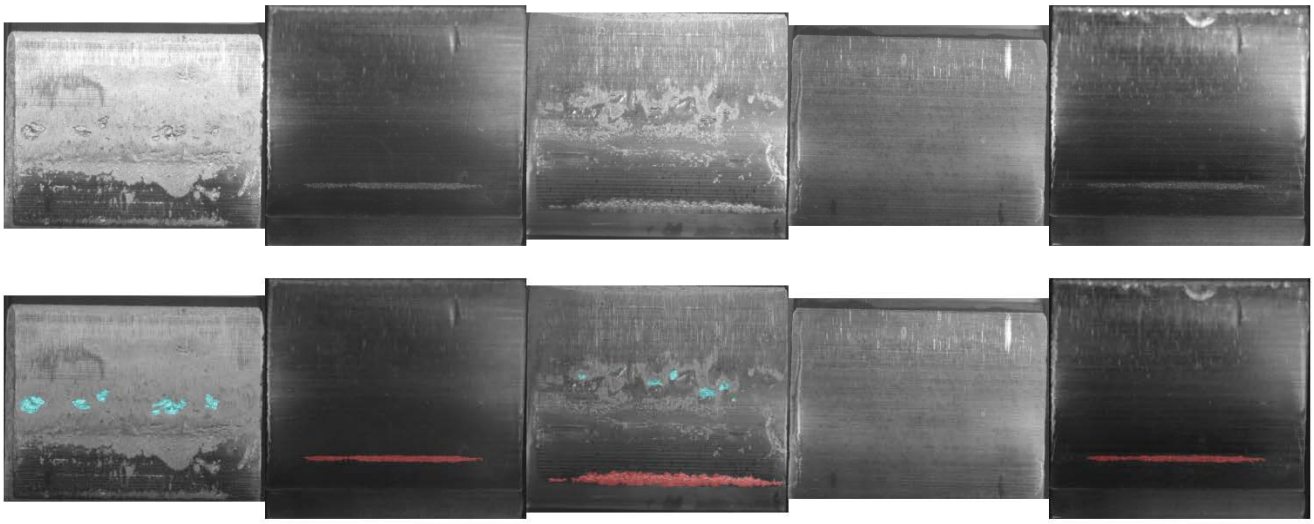


Figure 8. Uncertainties in image annotation. The first top 5 images (row 1) selected by active learning with BALD to be annotated, and their annotations (annotated by our experts in row 2).

ory, applications, and recent advances. *Expert Systems with Applications*, *221*, 119738.

Van Maele, D., Poletto, J. C., Neis, P., Ferreira, N., Fauconnier, D., & De Baets, P. (2023). Online vision-assisted condition monitoring of gearboxes. In *8th euro. conf. and exhibition on lubrication, maintenance and tribotech (lubmat 2023)*.

Wan, T., Xu, K., Yu, T., Wang, X., Feng, D., Ding, B., & Wang, H. (2023). A survey of deep active learning for foundation models. *Intelligent Computing*, *2*, 0058.

BIOGRAPHIES

Wenzhi Liao received the Ph.D. degree in Engineering from the South China University of Technology, Guangzhou, China, in 2012, and the Ph.D. degree in computer science engineering from Ghent University, Ghent, Belgium, in 2012. From 2012 to 2019, he has been working first as a Postdoc at Ghent University and then as a Postdoctoral Research Fellow for Research Foundation Flander (FWO). From February 2020 to January 2022, he had worked in VITO (Mol, Belgium) as a Data Scientist. Since February 2022, he works in Flanders Make, focusing on smart vision for Industry 4.0. His current research interests include Image Processing and Interpretation, Pattern Recognition, AI and Computer Vision. He is also highly experienced in Machine Learning, Large-scale problems and Remote Sensing. Dr. Liao was a recipient of the Best Paper Challenge Awards in both the 2013 IEEE GRSS Data Fusion Contest and the 2014 IEEE GRSS Data Fusion Contest. He serves as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS).

Roeland De Geest is a researcher at Flanders Make with 6+ years of experience in industry-driven research in the field of computer vision and machine learning. He obtained a Ph.D. degree in Electrical Engineering from the KU Leuven in 2019.

Djordy Van Maele Received his M. Sc. in electromechanical engineering technologies from Ghent University, Gent, Belgium, in 2021. In 2021 he started working on his doctoral degree in electromechanical engineering, in the field of tribology, at Ghent University until current date.

Jean Carlos Poletto obtained his M.Sc. degree in Mechanical Engineering from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2018. He has been developing research on experimental tribology, with active contributions to the field since 2015. Currently, he is working on a joint PhD program between UFRGS, Brazil, and Ghent University, Belgium.

Ted Ooijevaar is Senior Technology Domain Leader at Flanders Make with 8+ years of experience in industry-driven research and development for aerospace, machine and vehicle applications. Leads a research team focused on sensing (sensor fusion, virtual sensing), data analytics, modeling (physics, AI and physics augmented AI), monitoring (anomaly detection, diagnostics and prognostics) and health management technologies for machines and vehicles. Prior to this, he performed industry-driven research in the field of condition monitoring and data analytics as (senior) research engineer. He holds a Ph.D. degree in Mechanical Engineering from the University of Twente in the Netherlands in 2014. He also gained experience as a visiting Ph.D. researcher at the Nondestructive Evaluation Sciences Branch at the NASA Langley Research Center in the USA.

Laveen Prabhu Selvaraj is Senior Digital Solutions Engineer at ZF Wind Power Antwerpen NV from 2019. He is working on Condition Monitoring Systems (CMS) for Wind turbine gearbox, also working on new technology, sensor and innovation in WTG CMS systems. He was working on development of BIO sensors and characterization of Deep UV LEDS using electro luminescence spectroscopy (µEL) during his work as Scientific Researcher at Chemnitz University of Technology, Chemnitz, Germany for 3 years from 2016. He obtained his M.Sc. degree in Micro and Nano system from Chemnitz University of Technology, Chemnitz, Germany.

Luk Geens is a Senior Technology Engineer at ZF Wind Power Technology Antwerp (Belgium) with 20+ years of experience in wind turbine industry.

Advancing Durability Testing in Automotive Component through Prognostics and Health Management (PHM) Integration

Jinwoo Song[1], Junggyu Choi[1], Jeongmin Shin[1], Seungyoon Oh[1],
Seok Hyun Hong[2], Yun Jong Lee[2], Hae-Sung Yoon[1], and Joo-Ho Choi[1*]

[1]*Korea Aerospace University, Goyang-si, Gyeonggi-do, 10540, Republic of Korea*
jwsong@kau.kr
wndrb59@kau.kr
lgs360600@gmail.com
osyoony@kau.kr
hsyoon7@kau.ac.kr
**Corresponding author: jhchoi@kau.ac.kr*

[2]*Hyundai·Kia Motors R&D Center, Hwaseong-si, Gyeonggi-do, 18278, Republic of Korea*
ftrain77@genesis.com
yj.lee@genesis.com

ABSTRACT

In automotive Powered Door Systems (PDS), the emergence of grinding and clicking noise over time is a common failure mode. This issue typically arises from design or assembly inconsistencies and intensifies due to wear or increased clearance at its component, becoming noticeable to passengers, and causing discomfort. Numerous automotive manufacturers conduct comprehensive durability tests to tackle such issues during the development. Conventional durability tests, however, rely on the manual effort such as visual and auditory inspection at regular intervals, hence, is subjective and inefficient. This study introduces a novel method by the prognostics and health management (PHM) approach to detect anomaly and assess its severity of the noise during the durability test of the PDS, which may improve the reliability of noise detection and reduces the test time by early termination using prognosis capability. The results demonstrate the potential, paving the way for its broader application across various domains to advance testing processes and reliability.

1. INTRODUCTION

In recent developments in the automotive industry, many components are electrified to enhance the user convenience. Prominent examples include power window, automatic tailgates, and power door systems. These systems, however, often have various forms of wear and joint failures, significantly impacting user satisfaction and perceived vehicle quality.

Most issues with these components stem from design flaws or problems during the assembly process. To address these challenges and improve vehicle durability and reliability, automotive manufacturers conduct durability tests. Conventionally, these tests have relied on manual visual and auditory inspections performed at regular intervals, which are both subjective and inefficient.

Prognostics and Health Management (PHM) technology has gained considerable attention across various industries, including aerospace, smart manufacturing, power plants, and transportation, for its potential to prevent failures, reduce operational costs, and facilitate predictive maintenance.(Choi, 2014; Zio, 2022) The potential of PHM to enhance durability testing is substantial. In this paper, we discuss the application of PHM techniques and frameworks to durability testing, focusing on developing more accurate and automated diagnostic models.

1.1. Power Door System (PDS)

The case study presented in this paper focuses on the Power Door System (PDS), a feature designed to enhance user convenience in high-end vehicles. Figure 1 illustrates a vehicle equipped with the PDS on its rear door, showcasing the application of the PDS. This system automatically closes the door after a passenger enters.

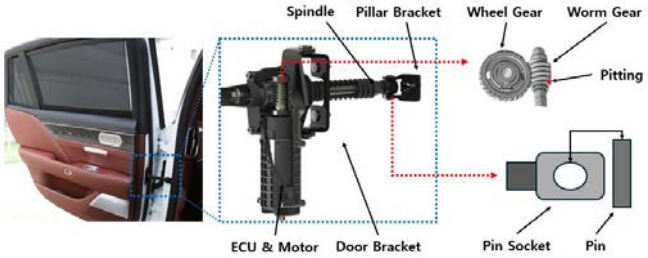


Figure 1. Power door system installed on a vehicle

The PDS is mounted on each door, and as shown in Figure 1, the ends of the spindles are connected into the pillar of vehicle. When the motor rotates, it drives the worm gear to turn the wheel gear. The relative motion then pulls the spindle, closing the door with a mechanism that takes about 3 seconds for the closure operation.

In case of the poor design or manufacturing, the PDS can encounter issues such as a grinding or clicking noises during the door closing, which usually occurs after cycles of door closing. The grinding noise is a continuous rough sound heard throughout the closure. This noise is often caused by small pitting on the drive worm gear, as depicted in upper right corner of Figure 1. On the other hand, the clicking noise is a sharp sound heard at a certain moment during the door closing. It occurs due to clearance, wear or assembly damage between the pin and pin socket connecting to the body's pillar. These noises are characterized as periodic and impulsive, respectively.

In order to determine the occurrence of these noises in the PDS, a durability test is conducted, incorporating both visual and auditory inspections at regular intervals. The method, however, is not reliable nor efficient due to the manual procedure. To overcome this, a diagnostic model is developed, thereby enhancing the reliability and functionality of the PDS.

1.2. Sensor Selection

To choose a sensor that can reveal useful features for the diagnosis, four sensors are considered for the potential candidates: motor current, motor rotation (hall sensor), and accelerometer attached to the PDS and to the body side. The current and hall sensor data are collected at a rate of 4 kHz, while the accelerometer data at 25.6 kHz.

Figure 2 illustrates the signals captured by these sensors during the operation of PDS. Each graph within the figure represents the data from different sensors, plotted over time to show the dynamic changes in sensor readings as the door progresses through 3 repetitions of open and close motions.

Upon comparison of these signals in terms of efficacy of diagnostic, consistency, strength and the convenience of installation, the accelerometer signals perpendicular to the body side is found appropriate for the study. Consequently, all the signals discussed in this paper are measured from this sensor.

The rest of the paper is outlined as follows: Section 2 introduces the PHM framework, Section 3 and 4 detail the process of developing diagnostic models for grinding noise and clicking noise, respectively. Lastly, conclusions are presented in Section 5.

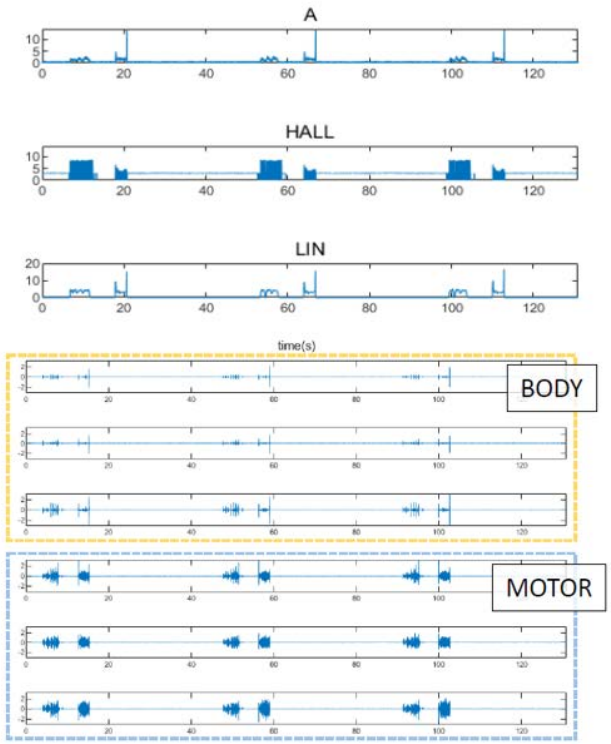


Figure 2. Various sensor signals of the PDS motion

2. PHM FRAMEWORK

In this section, the development procedure of diagnostic model and its application to test data are addressed. Figure 3 illustrates the overall flowchart of the PHM process. It has two major phases: construction of the diagnostic model and its application.

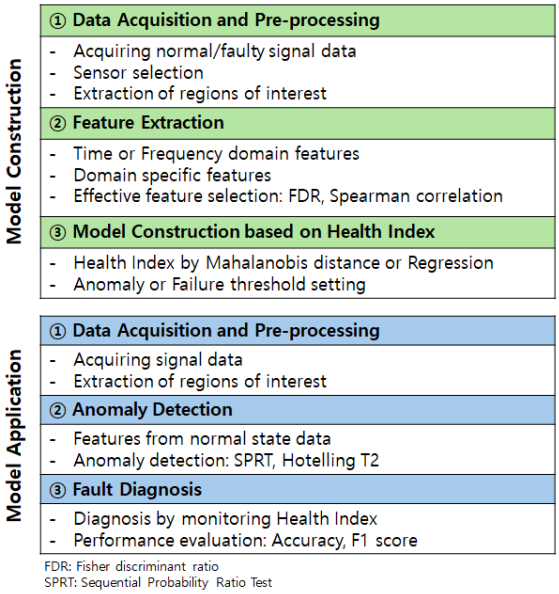


Figure 3. Flowchart of PHM framework

In the construction phase, the first step is the collection of sensor signals and pre-processing. Two types of signal data are collected in the test: first is the discrete data for normal and faulty conditions. second is the continuous data from the normal to the failure. In a single set of signal data, not the whole period is exploited but only a segment is taken for a better feature extraction. This varies depending on the considered noise, which accounts for the symptoms and causes of failures as well as the operational mechanisms of the system.

Next is the feature extraction and selection. Various features are extracted, including time domain, frequency domain, and domain-specific features.(Sim et al., 2020) Effective feature selection involves choosing the features with higher value of Fisher Discriminant Ratio (FDR) that can better distinguish the normal and faulty conditions in case of discrete data and the features with a higher Spearman correlation in case of the continuous data.

Based on the selected features, a health index (HI) is constructed in the next step. While there are several approaches for this, Mahalanobis distance is employed in this study, which is useful when there are the normal features only. Thresholds for anomaly and failure are established for the HI, respectively, to effectively distinguish between the normal, warning and failure states.

In the application phase, data from the test subjects are collected. Following the procedures defined in the diagnostic model, the HI is calculated and monitored to perform anomaly or fault detection. This systematic approach allows for precise and proactive management of system health.

3. Diagnosis Model for Grinding Noise

To develop the diagnostic model for grinding noise, normal and faulty data sets are collected from three specific cases. Among these, case A involves less severe noise occurrence, while the cases B and C involve relatively severe noise occurrences.

Table 1. grinding noise datasets

Vehicle / Placement	Class	Features	Case
Vehicle1 / Front	Normal	-	A
	Faulty	Small grinding noise	
Vehicle1 / Rear	Normal	-	B
	Faulty	Loud grinding noise	
Vehicle2 / Rear	Normal	-	C
	Faulty	Loud grinding noise	

As described in Section 2, for the development of the diagnostic model, accelerometer which attached to the vehicle's body is utilized. Figure 4 presents simultaneous recordings of the motor's relative rotational speed and vibration signals. It can be observed that the motor operates in three phases of acceleration, constant speed, and deceleration.

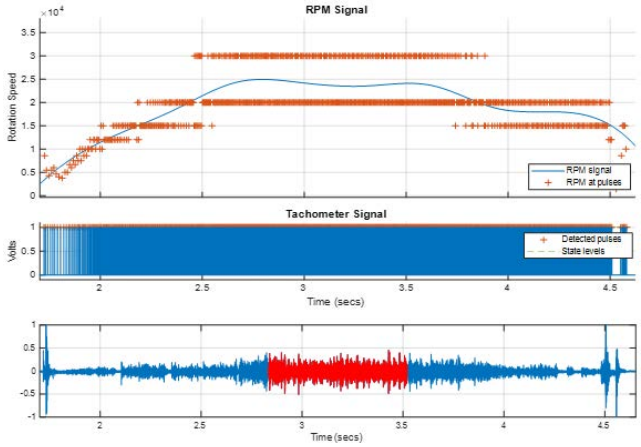


Figure 4. Motor speed and accelerometer signal while closing

Since the grinding noise occurs continuously during the closing operation, the signal over the whole period can be responsible for the noise. However, only a part with constant speed is chosen for the efficacy of feature extraction.

Using the signal in constant speed, numerous candidate features are extracted as shown in Figure 5, in which the blue o and red x denote the normal and fault respectively. Note that all the features are normalized by Gaussian distribution. Among these, more significant features are selected that can distinguish the two classes more clearly. For this purpose,

Fisher discriminant ratio (FDR) is calculated, defined as follows.

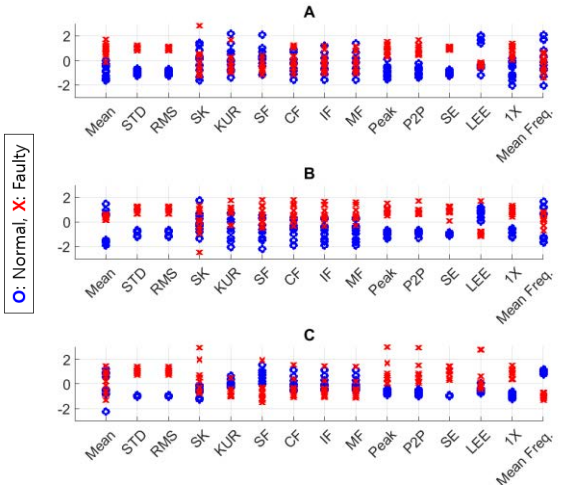


Figure 5. Features from each signal of four datasets

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (1)$$

where μ_i is mean value of feature, and σ_i is standard deviation of feature for i-th class. As a result, the root mean square (RMS) and Shannon entropy (SE) are recognized as the most important features.

Using these selected features, an HI based on the Mahalanobis distance is constructed from the normal data as defined in the following equation.

$$HI = (\mathbf{x} - \mu_n)' \mathbf{S}_n^{-1} (\mathbf{x} - \mu_n) \quad (2)$$

where $\mathbf{x}$ is feature vector of input data, μ_n is mean of feature vector for normal data and $\mathbf{S}_n$ is covariance matrix of feature for normal data. The results as shown in Figure 6, demonstrate a clear distinction between the normal and faulty data across all cases. Based on the HI values of the collected normal and faulty data, thresholds for anomaly and failure are established as the dotted blue and magenta lines, which are the normal HI at upper 95% confidence and fault HI at lower 95% confidence levels, respectively. This diagnostic model can be utilized in the future tests to determine whether the product yields the grinding noise during the cycles of operation.

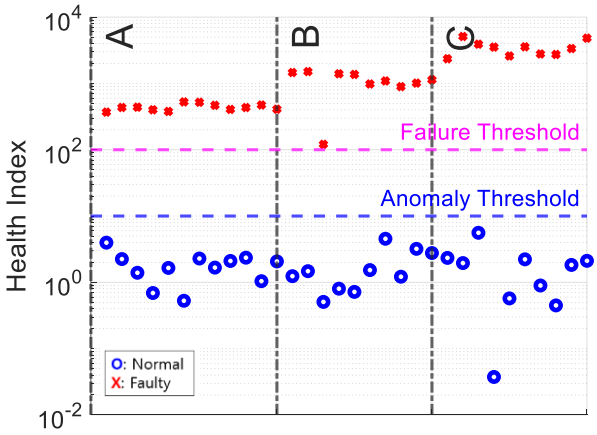


Figure 6. Health index for three cases

4. DIAGNOSIS MODEL FOR CLICKING NOISE

To develop the diagnostic model for clicking noise, datasets are collected from both the front and rear doors of vehicles. The collected datasets are given in Table 2. Front door datasets include four discrete states: two normals and two faults, each collected from different PDS installed on two vehicles. In the table, Normal 1 and Normal 2 indicate no clicking noise. However, Normal 2 is with a subtle grinding noise. And Fault 1 indicates a small clicking noise, while Fault 2 indicates a significantly loud clicking noise. For the rear door, run-to-failure data are collected for up to 38,000 cycles, from which the normal and fault are defined by those less than 10,000 and over 23,000 cycles based on the experts' opinion.

Table 2. Clicking noise datasets

Vehicle / Placement	Class	Feature	Case
Vehicle1 / Front	Normal1	-	Front
Vehicle1 / Front	Normal2	Small grinding noise from motor	
Vehicle2 / Front	Faulty1	Small clicking noise	
Vehicle1 / Front	Faulty2	Loud clicking noise	
Vehicle2 / Rear	Normal	- (Cycle 0 ~ 10000)	Rear
	Warning	Tiny clicking noise (Cycle 10000 ~ 23000)	

	Faulty	Small clicking noise (Cycle 23000 ~ 38000)	

As opposed to the grinding noise, clicking noise, characterized as an impulsive signal, typically occurs only at a certain moment during the door closing. To isolate these impulsive events, kurtosis in a short interval with 0.1 second is computed over a sliding time window, which is a widely used feature in vibration analysis for its ability to detect spikes in signals.(Cerrato-Jay et al., 2001; Honarvar & Martin, 1997)

Figure 7 illustrates both the original and its kurtosis for the sound and vibration signal, respectively. In the analysis, signals at the beginning (0-0.3 seconds) and after 2.5 seconds are disregarded as they are those at the start and end of closing, respectively. In the upper two figures, it is observed that the moments when the noise is heard and when its kurtosis shows local peak are the same, as marked by the red explosion symbols. Based on this finding, the vibration signal is processed in the similar manner, which are given in the lower two figures. Interestingly, the moments when the kurtosis shows local peak are the same in the sound and vibration signals. Therefore, the kurtosis is used as the means to identify the moment of clicking noise, and the signal over a small time period of 0.1 second is taken for further processing towards the feature extraction.

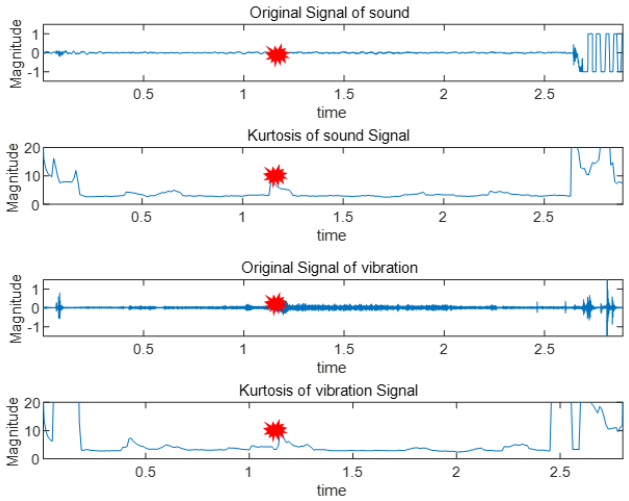
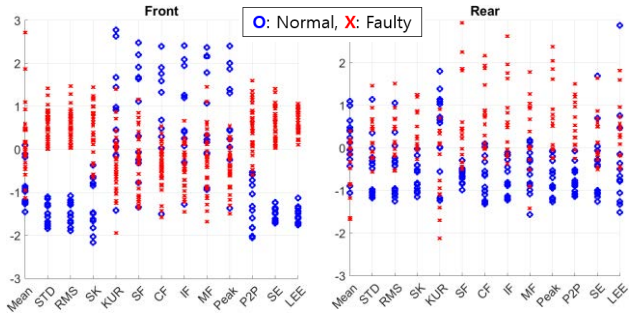


Figure 7. Raw signal and frame kurtosis for sound and vibration signal

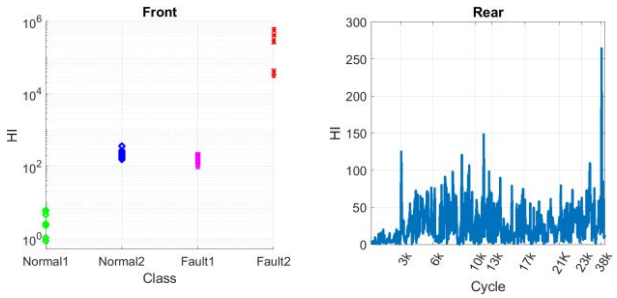
As in the previous section, several candidate features are extracted, from which the most significant ones are sought for. The results are in Figure 8, in which (a) are those for the front (normal 1 and fault 1 only), and (b) are for the rear (normal less than 3,000 cycles and fault over 33,000 cycles) are taken among the run-to-failure data). The blue o and red x denote the normal and fault respectively. In comparison with the grinding noise, the separation between the normal and fault both in the front and rear are less clear.



(a) (b)

Figure 8. Features from each dataset

Nevertheless, the same procedure is taken to select the most important features, which are RMS (root mean square), P2P (peak-to-peak), and SE (Shannon entropy). The diagnostic performance by the HI made of these features are shown in Figure 9. In Figure 9(a), which is the result of front door, considerable overlap is found between the HI for normal 2 (blue o) and fault 1 (magenta x). Furthermore, in the case of rear door as shown in Figure 9(b), clear increasing trend are not present towards the fault conditions. All these suggest that the selected features are not so effective to use to construct HI.



(a) (b)

Figure 9. Health Index from each dataset

To discover more effective features, another attempt is made, which is the time-frequency analysis. Figure 10 displays a spectrogram of continuous wavelet transformation (CWT), obtained for the instant of 0.1 second centered at the clicking noise. The result reveals that the clicking noise occurs in a very short 10 ms time window at a certain frequency range. In order to quantify this into the feature and use it as the HI, total energy of the impulsive moments within specific time and frequency windows is used. The results are illustrated in Figure 11, in which the Figure 11(a) for the front door shows a better distinction between the normal and fault, and Figure 11(b) for the rear door with run-to-failure presents a better increasing trend in HI, demonstrating the superiority of the CWT based approach over the time-based ones.

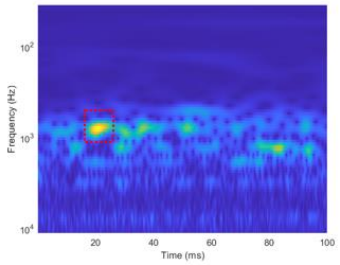
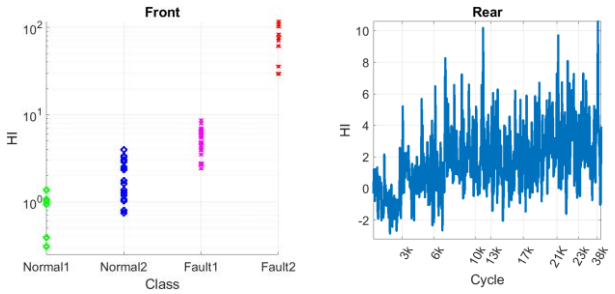


Figure 10. Spectrogram for click noise

Despite these advancements, there remains considerable dispersion in the HI, which makes it still challenging to apply in the field. This variability could stem from various external factors that influence the occurrence of clicking noises. Further signal processing efforts to mitigate these influences or the identification of more effective features will be essential to enhance diagnostic accuracy.



(a) (b)

Figure 11. Health Index from new feature

5. CONCLUSIONS

In this study, prognostics and health management (PHM) approach such as the signal processing, feature extraction and selection, and construction of HI, was applied to develop diagnostic models for two representative faults occurring in the power door systems (PDS): grinding and clicking noises. These faults are characterized by a continuous rough sound and a sharp, transient sound during door closure, respectively.

The method has facilitated the development of diagnostic models capable of detecting both types of noises, demonstrating the potential in the real-world applications.

However, much more data are necessary to refine the model and validate its performance, which requires a lot of efforts in time and money. Particularly for the clicking noise model, the HI contains significant uncertainty, highlighting the necessity for exploring diverse approaches and possibly new features to enhance diagnostic accuracy.

By integrating the PHM into the durability tests, we have showcased the potential for automation and quantitative fault assessment. If we can obtain comprehensive run-to-failure (RTF) data, it might also enable us to predict the remaining useful life of components, which could significantly reduce testing time by preemptively forecasting the occurrence of noise issues.

Acknowledgments:

This work was supported by NGV of Hyundai Motors company, which is greatly appreciated.

References:

Cerrato-Jay, G., Gabiniewicz, J., & Gatt, J. (2001). *Automatic Detection of Buzz, Squeak and Rattle Events DJ Pickering MTS Systems, Noise and Vibration Division*.

Choi, J.H. (2014). A review on prognostics and health management and its applications. *Journal of Aerospace System Engineering*, *8*(4), 7–17.

Honarvar, F., & Martin, H. R. (1997). *New Statistical Moments for Diagnostics of Rolling Element Bearings*. http://www.asme.org/about-asme/terms-of-use

Sim, J., Kim, S., Park, H. J., & Choi, J. H. (2020). A tutorial for feature engineering in the prognostics and health management of gears and bearings. *Applied Sciences (Switzerland)*, *10*(16). https://doi.org/10.3390/app10165639

Zio, E. (2022). Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering and System Safety*, *218*. https://doi.org/10.1016/j.ress.2021.108119

An Experiment on Anomaly Detection for Fault Vibration Signals Using Autoencoder-Based N-Segmentation Algorithm

Kichang Park[1] and Yongkwan Lee[2]

[1]*Intelligence Manufacturing Technology Institute, RESHENIE Co., Ltd., Suwon-si, Gyeonggi-do, 16229, Republic of Korea*
kc.park@reshenie.co.kr

[2]*Grand ICT R&D Center, Tech University of Korea, Siheung-si, Gyeonggi-do, 15073, Republic of Korea*
Ivan.lee@tukorea.ac.kr

ABSTRACT

Most manufacturing facilities driven by motors generate vibration and noise representing critical symptoms against facility malfunctioning conditions in the manufacturing industry. Due to the difficulty of obtaining abnormal data from facilities in manufacturing sites, many prior researchers who have studied predicting facility faults have adopted unsupervised learning-based anomaly detection approaches. Although these approaches have a strength requiring only data on from facility normal behaviors, it is not clear that the anomalies detected by an anomaly detection model are due to the real component faults. Also, the model performance is likely to change according to the diverse abnormal conditions of the given facility. In this paper, we took an experiment with a fault vibration simulator to measure the anomaly detection performance of a one-dimensional convolutional autoencoder model with different fault conditions. In the experiment, we used four different abnormal conditions: imbalance, misalignment, looseness, and bearing faults, which are the most frequently occurring facility component failures from the rotating machineries. Data were gathered from the simulator with the IEPE(Integrated Electronics Piezo-Electric) type sensor. We proposed the N-Segmentation algorithm that performs anomaly detection in segmented frequency region according to corresponding component faults for better anomaly detection performance. In conclusion, the proposed algorithm showed about 15 times better anomaly detection rate than not applying it.

1. INTRODUCTION

Artificial Intelligence (AI) technologies are adopted in various field domains to replace human beings or improve legacy systems. The manufacturing sector has also gradually tackled AI-based anomaly detection (AD) approaches for facility monitoring and fault detection. (Kumar, Khalid, & Kim, 2022; Zhang, Lin, Liu, Zhang, Yan, & Wei, 2019). AI-enabled facility monitoring systems are necessary to improve productivity, reduce costs, and ensure worker safety in manufacturing sites. Facility anomaly is an abnormal condition where defects or failures occur, and it can be determined and predicted by analyzing physical data measured during the facility operation from physical sensors such as those of vibration, current, and temperature. Since motors drive most manufacturing facilities, they generate various vibration signals during operation. These vibration signals represent a valuable basis for predicting whether the facility is in normal operations or defective status. When the vibration increases or becomes excessive, certain mechanical trouble has usually occurred. Since the vibration does not increase or become excessive for no reason, it is considered an indicator of machinery malfunction (Shreve, 1994). However, since the types of facility defects or failures are so diverse, obtaining sufficient data on facility defects and failures in manufacturing sites is impractical (Hiruta, Maki, Kato, & Umeda, 2021; Li, Li, & Ma, 2020). As a result, unsupervised learning approaches that use only data acquired when the facility is in a normal condition are very practical.

As the fault situations are diverse and unsupervised AD models require reconstruction of input signals, there are two important considerations, namely, types of faults (Thi, Do, Jung, Jo, & Kim, 2020) and the reconstruction range (Amarbayasgalan, Pham, Theera-Umpon, & Ryu, 2020). Considering these points, we conducted a vibration AD test using a fault vibration simulator and an Integrated Electronics Piezo-Electric (IEPE) type vibration sensor. We collected the vibration data of the simulator's normal signals and abnormal signals generated under normal operation and conditions of imbalance, misalignment, looseness, and bearing faults. Then, we trained the normal signals with a one-dimensional convolutional autoencoder (1-D CAE) and measured the AD performance by normal signals and fault signals. We introduce the N-segmentation algorithm for the better AD

performance, which performs it by segmenting the frequency range into N different regions. The proposed algorithm can detect fault vibration signals with improved performance.

2. BACKGROUND

AI models are classified as supervised or unsupervised learning, depending on whether labels are used during the learning process. Applying supervised learning to AD requires data for all types of anomalies. Since gathering anomaly data for enough training is practically impossible, unsupervised learning using only the normal-condition data of the facility is more suitable, and the Autoencoder (AE) is representative of unsupervised learning (Lee, Lee, & Kim, 2024; Wei, Jang-Jaccard, Xu, Sabrina, Camtepe, & Boulic, 2023). AE is an AI model that learns how to produce output data as close as possible to the input data without data labels. Using a difference between the input data of the AE and the reconstructed data generated by the output data of the AE detects anomalies. In this case, the reconstruction error, the difference between the input data and the output data of the AE, is calculated by error functions such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

Various research has been performed regarding AD using fault vibration signals. Wisal and Oh (2023) developed a new deep learning algorithm that utilized ResNet and convolutional neural networks to detect the unbalance of a rotating shaft for both binary and multiclass identification. Kamat et al. (2021) used random forest, artificial neural network, and AE to detect the bearing fault. Their experiment showed the AE provides the highest accuracy of 91% over the others. Ahmad et al. (2020) has taken experiment with AD for rotating machines by comparing a long short-term memory-based AE (LSTM-AE) and an isolation forest model. The experimental results on real-world datasets showed that the LSTM-AE achieved an average f1-score of 99.6%. Most previous works focused on developing the model achieving high accuracy for AD, but they are suggested within the limited fault environment or dataset.

In vibration accuracy for condition monitoring, an IEPE-type sensor is usually capable of more precise vibration measurement than a Micro-Elector-Mechanical System (MEMS) type sensor. (Hassan, Panduru, & Walsh, 2024). In the previous research regarding AD by frequency segmentation, Park & Lee (2022) successfully performed AD by synthesizing the frequency domain data of the IEPE-type vibration sensor collected from the printing facility and the virtual frequency signals. However, for the objective performance evaluation of the approach, it is necessary to utilize data collected in a simulator environment like actual facilities, not a virtual signal.

3. PROPOSED ALGORITHM

We propose the N-Segmentation algorithm that detects abnormal vibration signals by segmenting the frequency domain measured by a vibration sensor into N frequency ranges. The algorithm uses N reconstruction errors and N thresholds to determine anomalies in the target vibration signal. The algorithm predicts whether the frequency section corresponding to each segment is a normal or an anomaly using the threshold that is the maximum reconstruction error value of the segment. Therefore, the algorithm can perform not only AD of the target signal but also AD of the segments. In other words, it can present additional information on which segment of the entire frequency range has an anomaly occurred. Here, N, the number of segmented frequency ranges, is a kind of hyperparameter designated by analysts, so applying various N values is necessary to measure the performance of a model like 1-D CAE through the proposed algorithm. Figure 1 shows a schematic diagram of the proposed algorithm when N is 4.

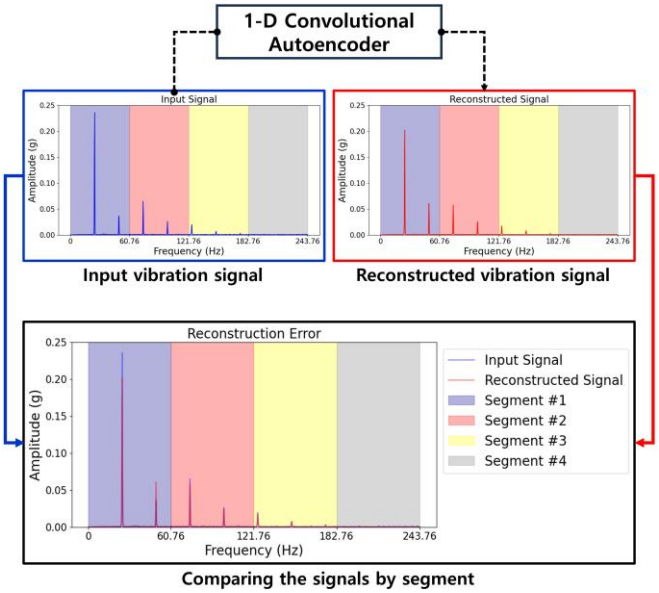


Figure 1. N-Segmentation Algorithm (N=4)

3.1. 1-Dimensional Convolutional Autoencoder

1-D CAE is an AE composed of encoders and decoders using one-dimensional convolutional layers (Zhang, Wang, Yi, Wang, Liu, & Chen, 2021). Convolutional layers can learn the data with high accuracy when the data size is constant, such as image data with two-dimensional data shape of width and height in pixels. Compared with two-dimensional images, as one-dimensional input vectors like vibration signals contain intuitive data characteristics, it is a great opportunity to use deep learning-based methodology for AD (Chen, Yu, & Wang, 2020). In the case of the frequency domain data used in this study, a 1-D CAE model is applied to capture the

status of the simulator with the vibration signal having one-dimensional 1024 numerical data. We developed the 1-D CAE model that consists of two 1-D convolution layers as encoder and two 1-D transposed layers as decoder using Python 3.10.9 and TensorFlow 1.12.0. Table 1 shows the structure of the model and hyperparameters.

Table 1. Structure of the model and hyperparameters

Layer	# of Filters	Kernel Size	Activation Function
1-D Conv.	64	8	RELU
1-D Conv.	32	8	RELU
1-D Trans. Conv.	64	8	RELU
1-D Trans. Conv.	1	8	-

Since MSE and RMSE can lead to higher weights given to higher errors, the model tends to be more sensitive to noise that might cause false positives (Kang, Kim, Kang, & Gwak, 2021). In the reconstruction loss function for AE-based AD models, MAE is more appropriate than the other two functions (Xu, Jang-Jaccard, Singh, Wei, & Sabrina, 2021). Therefore, the loss function used in the training process using the model is MAE in Eq. (1). Here, n denotes the number of numeric values contained in one vibration signal. X' denotes the reconstructed signal from the model, and X denotes the input signal to the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X'_i - X_i| \quad (1)$$

3.2. Thresholds

Unsupervised learning-based AD requires a threshold to determine whether the vibration signal represents an anomaly. In the existing AD using AE, the maximum reconstruction error value (Kang, Kim, Kang, & Gwak, 2021; Wei, Jang-Jaccard, Xu, Sabrina, Camtepe, & Boulic, 2023) or the 3-sigma value (Lee, Lee & Kim, 2024; Panza, Pota, & Esposito, 2023) among the reconstruction errors of the training data has been set as a threshold. In this case, the threshold has to be only one. However, in the proposed method, thresholds are generated as many as the number of segments N. Figure 2 shows the reconstruction error distributions as a histogram for each segment of the training data when N is 4. The red vertical line in each histogram in Figure 2 represents the maximum reconstruction error used as a threshold to determine the anomaly in the segment.

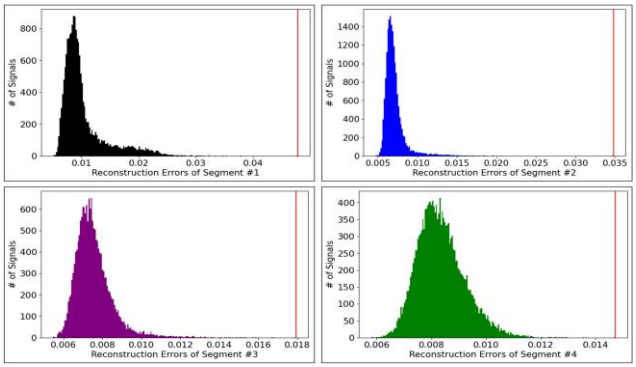


Figure 2. Reconstruction error distributions and thresholds (N=4)

3.3. Anomaly Detection

The proposed algorithm uses all N thresholds to detect the fault vibration signals among the test signals. If all the N reconstruction errors in the N segments of a target signal are lower than the corresponding thresholds, the signal is considered normal. Otherwise, it is a fault vibration signal. Figure 3 shows an example of the AD process of the proposed algorithm when N=2.

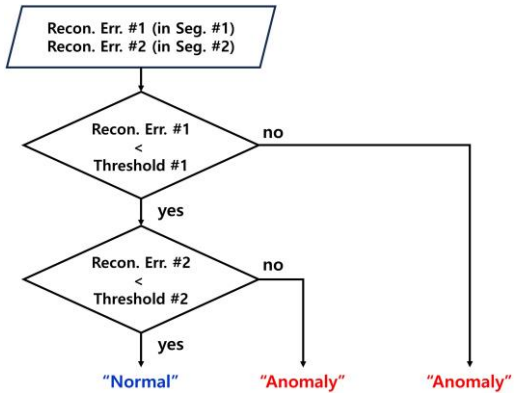


Figure 3. Process of anomaly detection (N=2)

The true-positive rate (TPR) in Eq. (2) was set to measure the performance of AD. Here, TP is the number of correctly predicted vibration signals, and FN is the number of incorrectly predicted ones.

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

4. EXPERIMENT

4.1. Setup

The vibration signals were gathered with a fault vibration simulator and an IEPE-type sensor for about two months. The simulator is the AST-VFS product of AST company of the Republic of Korea, and the sensor is the Model 131.02

product of VibraSens company of France. The vibration signals were collected every two minutes and obtained by Vib-AiR, the wireless health monitoring solution of RESHENIE company of the Republic of Korea, through open platform communication-unified architecture (OPC-UA) protocol (Schleipen, Gilani, Bischoff, & Pfrommer, 2016). Figure 4 shows the experimental environment. In Figure 4, R and L in parentheses mean Right and Left, respectively. The sensor (red square in Figure 4) was set on the top of the right-bearing housing of the simulator (blue square in Figure 4). The sensor can collect vibration signals in three-axis directions. In the experiment, only the Y-axis signals, which is the direction of rotation of the motor (the orange arrow in Figure 2), were used.

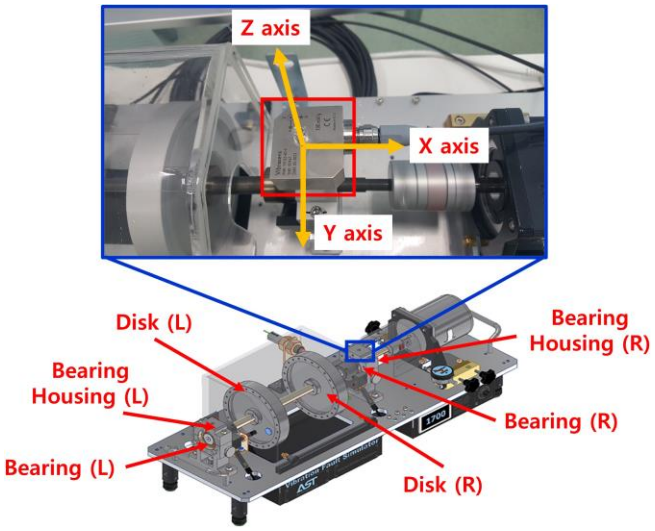


Figure 4. The fault vibration simulator and the IEPE sensor

4.2. Data Description

The dataset includes normal vibration signals at the motor rotation speed of 1,500 RPM and four fault vibration signals: imbalance, misalignment, looseness, and bearing faults. Table 2 summarizes the 16 simulator settings for generating fault signals. In Table 2, settings #1 to #6 are the conditions for the imbalance. Setting #7 is for the misalignment. Settings #8 to #10 are for the looseness. Finally, settings #11 to #16 are for the bearing faults.

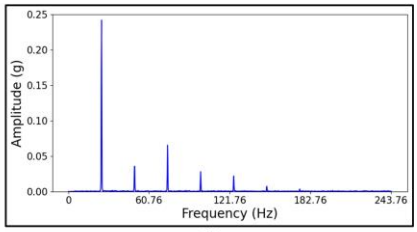
Table 2. Simulator settings for the faults

No	Setting
#1	Attaching 1.7g mass to the right disk
#2	Attaching 1.7g mass to the left disk
#3	Attaching 1.7g mass to each disk
#4	Attaching 6.25g mass to the right disk
#5	Attaching 6.25g mass to the left disk
#6	Attaching 6.25g mass to each disk
#7	Set to 1.2mm
#8	Loosening the left bearing housing
#9	Loosening the right bearing housing
#10	Loosening both bearing housing
#11	Applying outer wheel defect bearing to the left
#12	Applying outer wheel defect bearing to the right
#13	Applying inner wheel defect bearing to the left
#14	Applying inner wheel defect bearing to the right
#15	Applying ball defect bearing to the left
#16	Applying ball defect bearing to the right

One vibration data has 1,024 features that are numerical values representing amplitudes of each frequency from 0 to 243.76Hz. Figure 5 (a) is an example of the normal vibration data, and Figure 5 (b) is the result of visualizing it.

Frequency	Amplitude
0	0.000009
0.24	0.000015
0.71	0.000020
:	:
243.29	0.000458
243.52	0.000239
243.76	0.000327

(a)



(b)

Figure 5. An example of the normal data

Figure 6 shows the sample fault signals collected from the simulator using the sensor. Figure 6 (a), Figure 6 (b), Figure 6 (c), and Figure (d) are the results of the visualization of the signals collected under the imbalance of setting #1, the misalignment of setting #7, the looseness of setting #8, and the bearing fault of setting #11, respectively. These fault signals in Figure 6 showed different patterns in the number of peaks and amplitude values of peaks compared to the normal vibration signal in Figure 5(b). However, most frequencies in all signals showed very low amplitudes close to zero, except for a few frequencies. These frequency data characteristics affect the reconstruction results of the model and can consequently influence the AD performances.

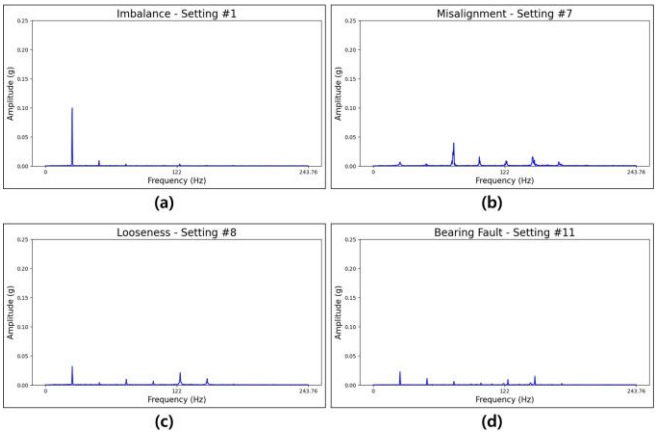


Figure 6. Example of the fault signals

A total of 21,365 vibration signals were collected from the simulator, of which 17,624 (82.5%) vibration signals were used as training data and 3,741 (17.5%) vibration signals as test data. Table 3 shows a detailed description of the dataset. Normal vibration signals of 17,624 training data and 481 test #1 data in Table 3 were generated when the motor in the simulator rotated at 1,500 RPM without applying fault condition settings described in Table 2. Test data used in tests #2 to #17 in Table 3 were generated from the operating conditions of the simulator with settings #1 to #16 in Table 2, respectively. In summary, the AD performances were measured with TPR for a total of 17 test cases (tests #1 to #17 in Table 3) in the experiment. The experiment was conducted in two cases: with and without the proposed algorithm.

Table 3. Description of the dataset

Purpose		Type	# of Data
Training		Normal	17,624
Test	#1	Normal	481
	#2	Imbalance (setting #1)	154
	#3	Imbalance (setting #2)	279
	#4	Imbalance (setting #3)	120
	#5	Imbalance (setting #4)	213
	#6	Imbalance (setting #5)	230
	#7	Imbalance (setting #6)	184
	#8	Misalignment (setting #7)	241
	#9	Looseness (setting #8)	319
	#10	Looseness (setting #9)	118
	#11	Looseness (setting #10)	265
	#12	Bearing Fault (setting #11)	138
	#13	Bearing Fault (setting #12)	251
	#14	Bearing Fault (setting #13)	127
	#15	Bearing Fault (setting #14)	269
	#16	Bearing Fault (setting #15)	196
	#17	Bearing Fault (setting #16)	156

5. Experimental Results

5.1. Anomaly Detection without N-Segmentation

Figure 7 shows the results of AD for the test data (Test #1 to #17 in Table 3) when the N-Segmentation algorithm is not applied. The blue dots and the red horizontal line in Figure 7 represent the MAE values for the test data and the threshold, respectively. Therefore, the dots above the threshold line mean predicted anomalies. The normal vibration signals (Test #1 in Figure 6) were exactly predicted as normal vibration signals. On the other hand, the performance of AD for the fault vibration signals (Test #2 to Test #17 in Figure 7) was too low. A few data were detected as anomaly signals in misalignment (Test #8 in Figure 7) and looseness (Test #10 and Test #11 in Figure 7), where TPRs were measured as 0.21, 0.19, and 0.05, respectively. Except for these cases, no detection was made for the rest of the anomalies. Overall, AD performances with the traditional approach using the model were too poor.

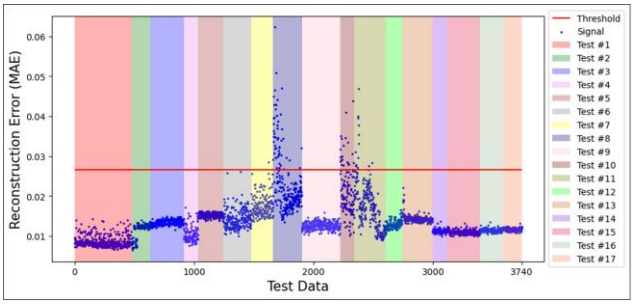


Figure 7. Result of anomaly detection without N-Segmentation

5.2. Anomaly Detection with N-Segmentation

As mentioned in Section 3.3, the proposed algorithm finally determines an anomaly signal by OR operation of AD results in segmented frequency ranges. Since the IEPE sensor collected a frequency signal of 0-243.76Hz, the segments are made by dividing the frequency range evenly. Figure 8 shows AD results in the segmented frequency ranges when N is 4: 0~60.76Hz (Figure 8 (a)), 61~121.76Hz (Figure 8 (b)), 122~182.76Hz (Figure 8 (c)), and 183.76Hz (Figure 8 (d)). In the example, as shown in Figure 8 (c) and Figure 8 (d), the proposed algorithm can detect most anomalies belonging to misalignment (Test #8 in Figure 8 (c)) and imbalance (Test #5 and Test #7 in Figure 8 (d)) compared to Figure 7. AD performance with the proposed algorithm has improved dramatically in these faults compared to the results of Figure 7.

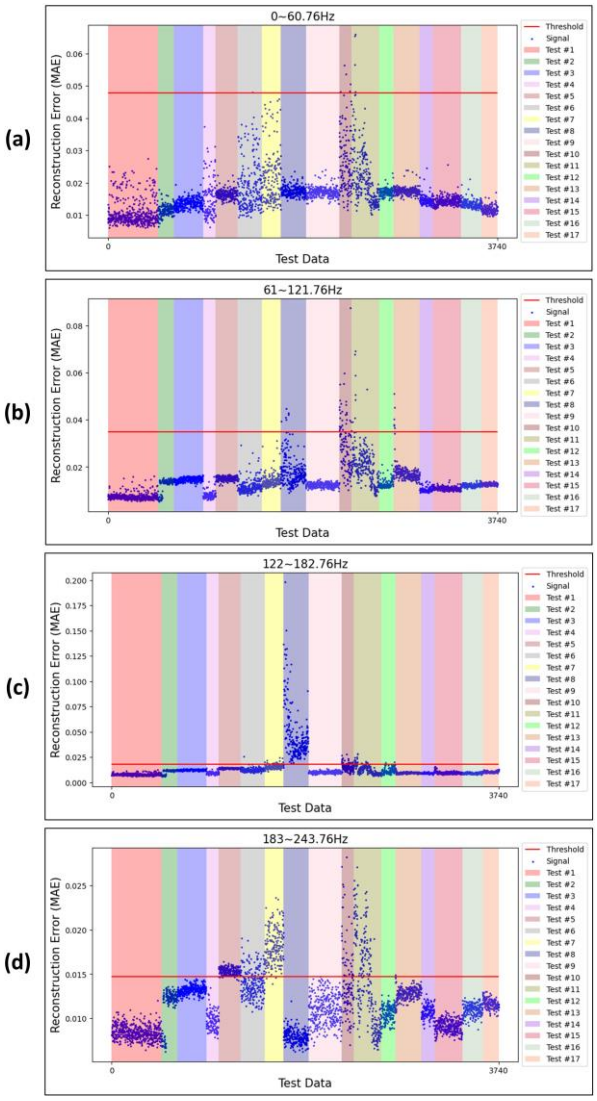


Figure 8. Result of anomaly detection with N-Segmentation (N=4)

As mentioned before, because N is a kind of hyperparameter, the proposed algorithm needs to be confirmed by changing N. Figure 9 shows the AD performance of the algorithm according to the change in the N value. When N is zero in Figure 9, the TPRs indicate AD results of not applying the proposed algorithm. The AD performance with N-Segmentation shows higher TPR scores than without it in all fault cases: imbalance (Figure 9 (a)), misalignment (Figure 9 (b)), Looseness (Figure 9 (c)), and Bearing Faults (Figure 9 (d)). Especially when N was 8, the TPRs for imbalance (Test #7) and misalignment (Test #8) improved dramatically from 0.01 to 0.99 and 0.21 to 1, respectively. In this case, the TPR for all fault signals was 0.40, and the proposed algorithm detected 1,301 anomalies among a total of 3,260 anomalies. On the other side, the traditional approach without the proposed algorithm just detected 87 anomalies.

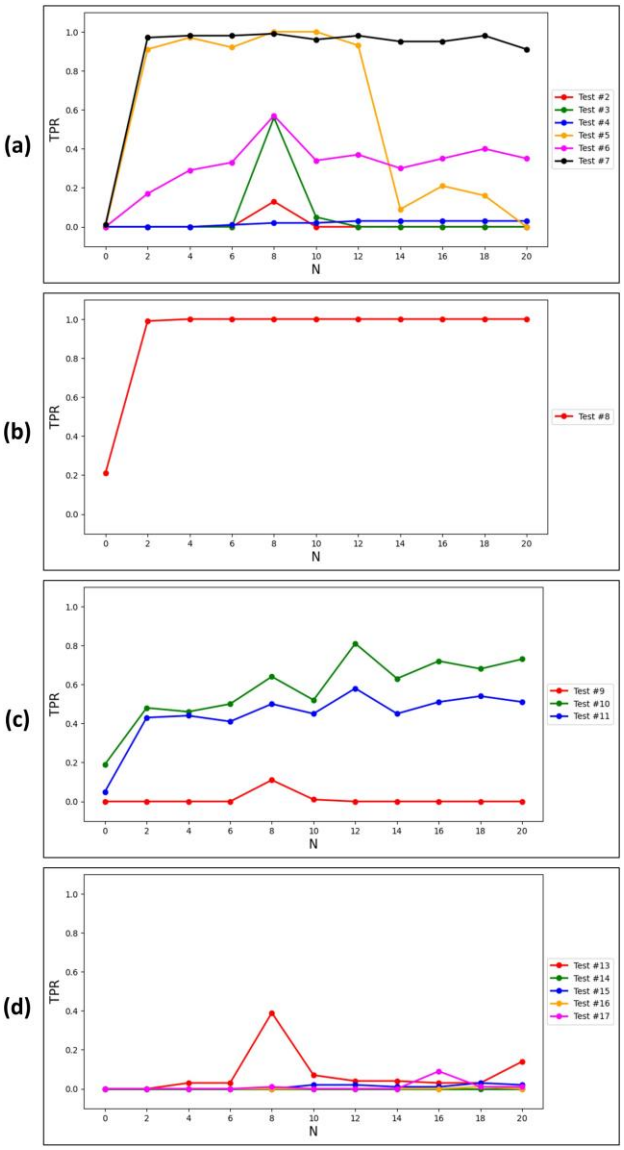


Figure 9. Performance of anomaly detection by N

6. Discussion

The results of the experiment can be summarized as follows. First, the N, the number of segmentations, did not affect the AD performance for the normal vibration signals in the test data. Regardless of N, the TPRs for the normal vibration signals were always measured as 1. Second, even if anomaly signals were the same fault type, the TPRs for the same fault signals showed differences according to the specific settings. Overall, anomaly signals measured in higher physical changes near the sensor were relatively better detected (see Table 2, Table 3, and Figure 9). Therefore, when detecting facility faults using a vibration sensor, the position of the sensor must be carefully determined. Third, the proposed algorithm can improve the performance of the unsupervised-based AD. In our experiment, the proposed algorithm

detected about 15 times fault vibration signals in the best case (N=8) than N=0. Even in the worst case (N=14), it could detect fault vibration signals more than 8 times. Fourth, the proposed algorithm not only detects the fault vibration signals with better performance but also provides additional information about the frequency range in which the anomaly occurred (see Figure 8). This information can be used to predict the type of facility failure.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the N-segmentation algorithm, which uses segmented frequency ranges to enhance facility fault detection performance. To measure the algorithm performance, we collected the frequency domain data of the vibration signal with the fault vibration simulator using the IEPE-type sensor. We trained the normal vibration data in the collected vibration signal using the 1-D CAE model and performed AD for the normal vibration data and four types of fault vibration data: imbalance, misalignment, looseness, and bearing faults. We detected up to 15 times more anomalies with the proposed algorithm than without it. The results show that the proposed algorithm is effective in AD for fault vibration signals. However, this study has limitations in applying only the 1-D CAE model and experimenting in the simulation environment. In future research, we aim to improve the proposed algorithm by comparing other machine learning models, and we will adopt it to facilities and equipment operating in real manufacturing sites.

ACKNOWLEDGEMENT

This work was partly supported by Innovative Hunam Resource Development for Local Intellectualization program through the Institute of Information & Communication Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2024-2020-0-01741, 50%) and partly supported by the project for Smart Manufacturing Innovation R&D funded Korea Ministry of SMEs and Startups in 2022. (Project No. RS-2022-00141076, 50%)

REFERENCES

Ahmad, S., Styp-Rekowski, K., Nedelkoski, S., & Kao, O. (2020). Autoencoder-based condition monitoring and anomaly detection method for rotating machines. *In 2020 IEEE International Conference on Big Data (Big Data)* (pp. 4093-4102). IEEE.

Amarbayasgalan, T., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (2020). Unsupervised anomaly detection approach for time-series in multi-domains using deep reconstruction error. *Symmetry*, 12(8), 1251.

Chen, S., Yu, J., & Wang, S. (2020). One-dimensional convolutional auto-encoder-based feature learning for fault diagnosis of multivariate processes. *Journal of Process Control*, 87, 54-67.

Hassan, I. U., Panduru, K., & Walsh, J. (2024). An In-Depth Study of Vibration Sensors for Condition Monitoring. *Sensors*, 24(3), 740.

Hiruta, T., Maki, K., Kato, T., & Umeda, Y. (2021). Unsupervised learning based diagnosis model for anomaly detection of motor bearing with current data. *Procedia CIRP*, 98, 336-341.

Kamat, P., Marni, P., Cardoz, L., Irani, A., Gajula, A., Saha, A., Kumar, S. & Sugandhi, R. (2021). Bearing fault detection using comparative analysis of random forest, ANN, and autoencoder methods. *In Communication and Intelligent Systems: Proceedings of ICCIS 2020* (pp. 157-171). Springer Singapore.

Kang, J., Kim, C. S., Kang, J. W., & Gwak, J. (2021). Anomaly detection of the brake operating unit on metro vehicles using a one-class LSTM autoencoder. *Applied Sciences*, 11(19), 9290.

Kumar, P., Khalid, S., & Kim, H. S. (2023). Prognostics and Health Management of Rotating Machinery of Industrial Robot with Deep Learning Applications—A Review. *Mathematics*, 11(13), 3008.

Lee, Y. K., Lee, S., & Kim, S. H. (2024). Real-Time Defect Monitoring of Laser Micro-drilling Using Reflective Light and Machine Learning Models. *International Journal of Precision Engineering and Manufacturing*, 25(1), 155-164.

Li, X., Li, X., & Ma, H. (2020). Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery. *Mechanical Systems and Signal Processing*, 143, 106825.

Panza M. A., Pota M., & Esposito, M. (2023). Anomaly Detection Methods for Industrial Applications: A Comparative Study. *Electronics*. 2023; 12(18)

Park, K., & Lee, Y. (2023). Anomaly Detection in a Combined Driving System based on Unsupervised Learning. *Journal of the Korean Society for Precision Engineering*, 40(11), 921-928.

Schleipen, M., Gilani, S. S., Bischoff, T., & Pfrommer, J. (2016). OPC UA & Industrie 4.0-enabling technology with high diversity and variability. *Procedia Cirp*, 57, 315-320.

Shreve, D. H. (1994). Introduction to vibration technology. *Proceedings of Predictive Maintenance Technology Conference*. November.

Thi, N. D. T., Do, T. D., Jung, J. R., Jo, H., & Kim, Y. H. (2020). Anomaly detection for partial discharge in gas-insulated switchgears using autoencoder. *IEEE Access*, 8, 152248-152257.

Wei, Y., Jang-Jaccard, J., Xu, W., Sabrina, F., Camtepe, S., & Boulic, M. (2023). LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. *IEEE Sensors Journal*, 23(4), 3787-3800.

Wisal, M., & Oh, K. Y. (2023). A New Deep Learning Framework for Imbalance Detection of a Rotating Shaft. *Sensors*, 23(16), 7141.

Xu, W., Jang-Jaccard, J., Singh, A., Wei, Y., & Sabrina, F. (2021). Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset. *IEEE Access*, 9, 140136-140146.

Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., & Wei, M. (2019). A review on deep learning applications in prognostics and health management. *IEEE Access*, 7, 162415-162438.

Zhang, Y., Wang, Y., Yi, Y., Wang, J., Liu, J., & Chen, Z. (2021). Coupling matrix extraction of microwave filters by using one-dimensional convolutional autoencoders. *Frontiers in Physics*, 521.

BIOGRAPHIES

Kichang Park received his M.S. and Ph.D. in Computer Science from Chonnam National University, the Republic of Korea, in 2003 and 2013, respectively. He has a background as a software engineer in the manufacturing industry and is currently working as an artificial intelligence professional at the Intelligent Manufacturing Technology Institute, RESHENIE Co. Ltd. His current research interest is AI-based anomaly detection for manufacturing facilities.

Yongkwan Lee is a Professor at the Tech University of Korea and founder of manufacturing AI solution company, RESHENIE Co. Ltd. He finished Bachelor's and Master's course at Kumoh National University of Technology, the Republic of Korea and got a Ph.D. degree in St. Petersburg State Polytechnic University, Russia. His research interest is manufacturing artificial intelligence, facility diagnosis technology, and smart-manufacturing solutions.

Analytical Modeling of Health Indices for Prognostics and Health Management

Pierre Dersin[1], Kristupas Bajarunas[2,3], and Manuel Arias Chao[2,3]

[1] *Luleå University of Technology, Luleå, 97187, Sweden*
pierre.dersin@ltu.se

[2] *Delft University of Technology, 292F+VC Delft, Netherlands*
k.v.b.bajarunas@tudelft.nl
m.a.c.ariaschao@tudelft.nl

[3] *Zurich University of Applied Sciences, 8400 Winterthur, Switzerland*
baja@zhaw.ch
aria@zhaw.ch

ABSTRACT

Understanding the current health condition of complex systems and their temporal evolution is an important step in prognostics and health management (PHM). However, when managing a fleet of complex systems, variations arising from manufacturing, environmental factors, mission profiles, and maintenance practices result in diverse health index (HI) trajectories. Therefore, in PHM, it is essential not only to identify common fleet-wide trends but also to account for individual asset-level variations when inferring HIs.

While several data-driven approaches exist for inferring individual asset-level HIs from unsupervised run-to-failure degradation data (see e.g. (Djeziri, Benmoussa, & Zio, 2020)), little research has been devoted to deriving analytical probabilistic representations of HIs that encompass both fleet-level trends and individual asset-level fluctuations. This paper aims to bridge this gap by addressing the research question of how to obtain an analytical representation of probability distributions for the time to reach intermediate degradation levels, using run-to-failure data or incomplete degradation trajectories from a fleet of complex systems.

In this work, it is assumed that suitable, asset-specific HI curves have been inferred through a fusion of deep learning techniques and prior expert knowledge of degradation physics (e.g., (Bajarunas, Baptista, Goebel, & Chao, 2023)). Given this context, we derive an analytical probabilistic description of the health index (HI) that reflects both fleet-wide trends and asset-specific conditions in the cases of Gamma or Weibull time-to-failure (TTF) distributions. Our approach involves defining HIs with a power law function, enabling the modeling of TTF and time to reach intermediate degradation levels. Moreover, we also detail the procedure for estimating the power law exponent from field data through regression analysis and conduct a sensitivity analysis regarding this exponent.

To illustrate our methodology, we present two case studies based on the N-CMAPPS dataset of turbofan engines and Li-ion batteries, validating the aforementioned assumptions and illustrating our methodology steps.

1. INTRODUCTION

An important step in prognostics and health management of complex industrial systems is inferring their current health condition. To this end, a normalized health index is often defined as a metric that measures the degree of degradation of equipment. Conventionally, a value of 1 for the health index corresponds to perfect health, and a value of 0 to a failed state. An intermediate value characterizes a state where the item is still operating but less than perfectly. If the health index captures the physical condition of the asset correctly, the time evolution of the health index is an appropriate means for performing prognostics, i.e., predicting the evolution of a degradation, eventually up to a failure, and the time until that failure, or remaining useful life (RUL). Therefore, the health index for an asset constitutes a key tool for maintenance decision-making, as it enables health assessment (in particular, degradation severity) and prognostics.

The derivation of HIs has traditionally depended on extracting key features from condition monitoring (CM) data and integrating them with a physical understanding of the asset to create a health index (Atamuradov et al., 2020). This practice, while effective, is heavily reliant on domain-specific knowledge, presenting a significant barrier to scalability and adaptability across different systems. To address these limitations, diverse data-driven approaches have been proposed for estimating HI from condition monitoring data. For instance, supervised learning models have been applied when dealing with datasets that contain labels of HIs (Roman, Saxena, Robu, Pecht, & Flynn, 2021). Similarly, residual techniques that identify deviations from a system's expected behavior (Ye & Yu, 2021; Hsu, Frusque, & Fink, 2023) offer another pathway, albeit contingent on the existence of a representative dataset of "healthy" state labels- an often challenging prerequisite in industrial settings due to difficulties in obtaining a representative data for complex systems. Recently, unsupervised methods combining deep learning methodologies with traditional reliability engineering principles in the form of explicit, analytical representation of the health index have shown promise in inferring asset-specific HI (Bajarunas et al., 2023; Yang, Habibullah, & Shen, 2021; Qin, Yang, Zhou, Pu, & Mao, 2023). Therefore, these recent works highlight the potential of leveraging the extensive body of reliability engineering theory, alongside deep learning algorithms, to model RUL dynamics effectively. An in-depth study of RUL dynamics and uncertainty, based on reliability theory, is reported in (Dersin, 2023).

In this work, our objective is to provide a theoretical foundation for constructing a robust analytical HI that reflects both fleet-wide trends and asset-specific conditions. By doing this, we aim to enable the integration of reliability engineering models in machine learning algorithms by providing an analytical probabilistic description of the HI. Addressing this objective involves answering the following question: *How to find an analytical description for a time-dependent health index integrating random parameters to capture asset variability and align with observed times to reach various degradation severity levels including the time to failure?*

Hence, in this work, we assume the availability of time-to-failure (TTF) distributions for a fleet of assets. Given this assumption, we formulate the problem in a general context and provide an analytical solution when the TTF follows a Gamma distribution or a Weibull distribution. In this scenario, with a health index defined by a power law featuring either an inverse-Gamma or a Fréchet-distributed coefficient, as the case may be, we demonstrate that the time to reach any intermediate degradation level follows a Gamma or Weibull distribution, respectively, sharing the same shape parameter as the TTF. Moreover, the scale parameter explicitly depends on the degradation level. We also detail the procedure for estimating the power law exponent from field data through regression analysis and conduct a sensitivity analysis regarding this exponent.

To validate our methodology, we present case studies focusing on the N-CMAPPS turbofan and randomized usage Li-ion batteries datasets. The results confirm the proposed methodology and highlights its practical applications. Obtaining an explicit, analytical representation of the health index, including the random variability among assets, is a definite advance over the state of the art that offers a major advantage. The proposed approach enables maintenance decision-making with minimal computational demand.

The paper is organized as follows: Section 2 presents the methodology used in this work; we first formulate the problem in Section 2.1 and present a resolution method in Section 2.2. We then delve into specific cases involving Gamma (Section 2.3) and Weibull distributions (Section 2.4), followed by a discussion on estimating the power law exponent controlling the shape of degradation for both analyzed distributions (Section 2.5) . Case studies from the N-CMAPSS and randomized battery usage datasets illustrate our approach (Section 3), with sensitivity analysis on the power law exponent (Section 4). The paper concludes with a summary of our findings and suggestions for future research in Section 5.

2. METHOD

This section provides a detailed explanation of the methodology used to derive an analytical description of the HI. The process is divided into several steps, which are outlined below and visually summarized in Figure 1

2.1. Problem Statement

A degradation phenomenon can be described by an HI, which evolves with time i.e., $HI(t)$, usually monotonically, from a perfect health condition to a failed state. Perfect health corresponds to a value $HI(t) = 1$, and failure is deemed to occur at the first time t, for which $HI(t)$ hits 0.

Given a plausible probability distribution for the time to failure, denoted T, which is derived from available data or prior knowledge, it is desired to find a family of probability distributions for the times T_s needed for the HI to reach any intermediate health level s,

$$0 < s < 1 \tag{1}$$

In other words, given a prior probability distribution, conditional upon $HI(0) = 1$, for the time to failure T,

$$T = \inf[t : HI(t) = 0] \tag{2}$$

find, for any intermediate level s, the probability distribution for T_s:

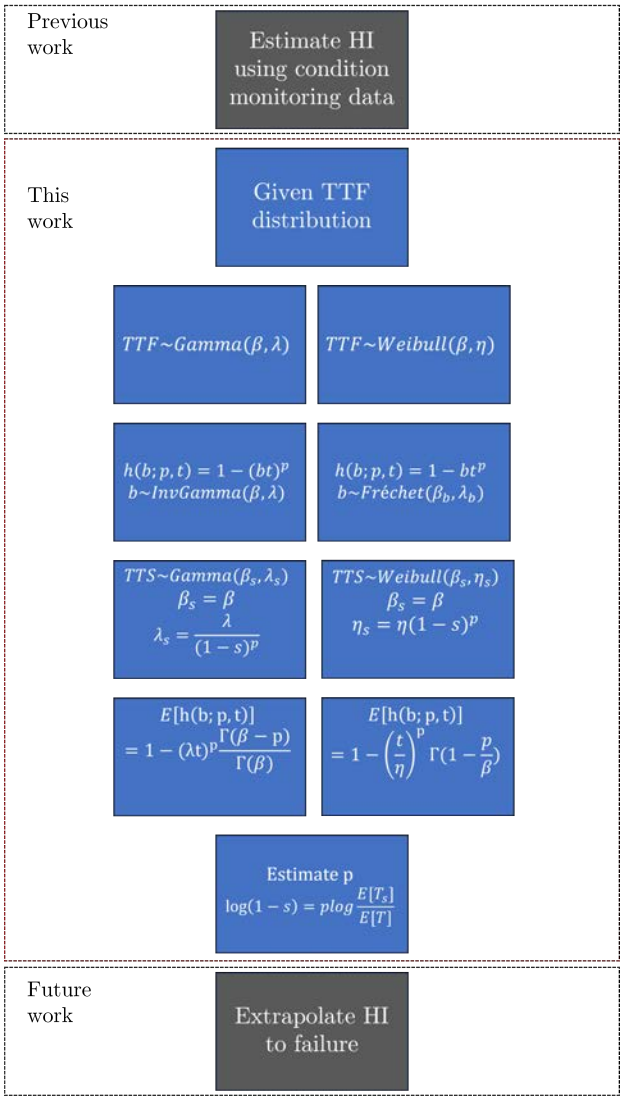


Figure 1. Flowchart illustrating the different steps of the proposed method along with the context of its applicability. (*Previous work*) The methodology assumes the availability of asset-specific Health Index (HI) derived from CM data, for instance, based on Bajarunas et al. (*This work*) The process of estimating begins with the assumption of a prior plausible Time to Failure (TTF) distribution (Gamma or Weibull models) and a probabilistic parametric model of a Health Index ($h(b, p, t)$). Based on these assumptions, we derive the analytical forms of the distributions describing the time required for the HI to reach any specified intermediate health level (TTS), thus providing a comprehensive statistical framework to model a degradation process. (*Future work*) The estimation of probabilistic, asset-specific failure times through extrapolation of individual HI's is suggested as a possible application of the derived analytical HI.

2.2. Resolution Method: General Principle

Let $R(t)$ denote an assumed reliability function. Then a probabilistic model for $HI(t)$, as a non-increasing function of t, is selected, and the condition $P[T > t] = R(t)$ is imposed. Finally, Eq. (3) is applied to obtain the distribution of T_s:

$$R_s(t) = P[T_s > t] \tag{4}$$

Let us consider the following parametric model for the health index:

$$HI(t) = h(p_1, p_2, ...p_n; t) \tag{5}$$

with an assumed functional form h, where some of the parameters $p_1, p_2, \ldots, p_n$ are random variables.

Then, it should be noted that

$$h(p_1, p_2, ...p_n; t) > 0 \tag{6}$$

is equivalent to

$$T > t \tag{7}$$

therefore the following condition is imposed:

$$P[h(p_1, p_2, ...p_n; t) > 0] = R(t) \tag{8}$$

with the right-hand side of Eq. (8) known.

Similarly, the condition $T_s > t$ is equivalent to $HI(t) > s$ and hence from Eq. (8), one derives

$$P[h(p_1, p_2, ..., p_n; t) > s] = R_s(t) \tag{9}$$

for any value of s between 0 and 1.

The method is quite general and can be applied to any TTF distribution. In the next two subsections, the method is detailed and illustrated on two frequently encountered families of TTF distributions: Gamma and Weibull, respectively.

2.3. Gamma Case

Let us consider the case when the time to failure follows a Gamma distribution with shape parameter β and rate parameter λ. The Gamma reliability function for time T (Nachlas, 2017) can be expressed as:

$$R(t) = 1 - \frac{\gamma(\lambda t; \beta)}{\Gamma(\beta)} \tag{10}$$

where $\gamma(\lambda.t; \beta)$ stands for the incomplete Euler gamma function.

A health index is sought, $HI(t)$, such that the time for the HI to reach the value 0 is Gamma-distributed.

We shall now show that a solution is provided by the following power law for the health index:

$$T_s = \inf[t : HI(t) = s] \tag{3}$$

$$h(b; p; t) = 1 - (bt)^p \tag{11}$$

with a positive exponent p and a random variable b with an inverse-gamma distribution with shape parameter β and scale parameter λ (b has the dimension of a frequency, i.e., the inverse of a time, so does λ). The health index defined by Equation (11) decreases monotonically from 1 to 0 as the time or usage variable t increases from 0 to $\frac{1}{b}$. It is a convex function of t if $p < 1$ and a concave function if $p > 1$ (and linear in the limit case of $p = 1$). The property that b has an inverse-Gamma distribution is equivalent to $\frac{1}{b}$ having a Gamma distribution with parameters β (shape) and λ (rate).

Denoting by T the time to failure, there follows from the above health index definition that

$$P[T > t] = P[(bt)^p < 1] = P[bt < 1] = P[\frac{1}{b} > t] \quad (12)$$

Since $\frac{1}{b}$ is Gamma distributed , the right-hand side of (12) is the Gamma reliability function at time t, with shape and rate parameters respectively equal to β and λ. Therefore, it has been proved that the definition (11) for the health index leads to a Gamma-distributed time to failure. .

Now let us look at the distribution of the time for the health index to reach a level s, between 0 and 1.

Let us denote that first hitting time T_s.

$$P[T_s > t] = P[h(b; p; t) > s] = P[1 - (bt)^p > s] \quad (13)$$

Equation (13) is equivalent to:

$$P[T_s > t] = P[(bt)^p < 1 - s] = P[\frac{1}{b} > \frac{t}{(1-s)^{\frac{1}{p}}}] \quad (14)$$

Since $\frac{1}{b}$ is Gamma (β,λ) distributed, it follows from (10) that,

$$R_{T_s}(t) = P[T_s > t] = 1 - \frac{\gamma(\frac{\lambda t}{(1-s)^{\frac{1}{p}}}; \beta)}{\Gamma(\beta)} \quad (15)$$

Therefore it has been shown that T_s has a Gamma distribution with shape factor β, and rate parameter λ_s given by the following function of s and the exponent p:

$$\lambda_s = \frac{\lambda}{(1-s)^{\frac{1}{p}}} \quad (16)$$

The problem stated in the beginning has thus been solved in the case when the time to failure has a Gamma distribution. The mathematical expectations of T_s and that of the health index $HI(t)$ are then derived explicitly, as follows, from the properties of the gamma distribution and the inverse-gamma distribution (Llera & Beckmann, 2016):

$$E(T_s) = \frac{\beta}{\lambda_s} = \frac{\beta}{\lambda}(1-s)^{\frac{1}{p}} \quad (17)$$

which can also be written as :

$$E(T_s) = E(T)(1-s)^{\frac{1}{p}} \quad (18)$$

To derive the expectation of the health index $HI(t)$; we now use properties of the inverse-gamma distribution. If X has an inverse-gamma distribution with parameters β and λ, the nth-order moment of X is given (Llera & Beckmann, 2016) by:

$$E(X^n) = \lambda^n \frac{\Gamma(\beta - n)}{\Gamma(\beta)} \quad (19)$$

as long as

$$n < \beta$$

Therefore

$$E[HI(t)] = 1 - E(b^p)t^p = 1 - (\lambda t)^p \frac{\Gamma(\beta - p)}{\Gamma(\beta)} \quad (20)$$

assuming the exponent p to be smaller than the shape factor β.

2.4. Weibull Case

We shall now consider the case where the time to failure follows a 2-parameter Weibull distribution. Denoting β and η the shape and scale parameters, respectively, this corresponds to the well-know reliability function:

$$R(t) = e^{-(t/\eta)^\beta} \quad (21)$$

For the health index, let us take the following power law, slightly different from the one taken in the Gamma distribution case, for reasons which will become apparent:

$$h(b; p; t) = 1 - bt^p \quad (22)$$

where p is a positive exponent, and b is a random variable. It will be seen that, if b has a Fréchet distribution (Fréchet, 1927; Ramos, Louzada, Ramos, & Dey, 2020), then the time to failure is Weibull distributed.

Indeed, by definition of the Fréchet (also known as "inverse Weibull") distribution, if the random variable b is Fréchet-distributed with scale parameter λ_b and shape parameter β_b, then:

$$P[b > u] = 1 - exp[-(\frac{u}{\lambda_b})^{-\beta_b})] \quad (23)$$

The shape parameter β_b is dimensionless, and the scale parameter λ_b has the dimension of t to the power of $(-p)$, just as the coefficient b.

Then, by substituting

$$u = t^{-p} \quad (24)$$

in (23), the following is obtained :

$$P[HI(t) > 0] = P[b < t^{-p}] = exp[-(\lambda_b t^p)^{\beta_b}] \quad (25)$$

and this expression must be equated to $P[T > t]$, which is assumed to be the reliability function of a two-parameter Weibull variable (η ,β).

Therefore, the parameters of the Fréchet distribution for b are obtained as follows:

$$\lambda_b = 1/\eta^p \quad (26)$$

$$\beta_b = \beta/p \quad (27)$$

as it can be verified by substituting the right-hand sides of (26) and (27) respectively for λ_b and β_b in (25). Then the distribution of T_s, the first hitting time of level s, can be derived as well, for any value of s between 0 and 1.

$$P[Ts > t] = P[HI(t) > s] \quad (28)$$

$$= P[1 - bt^p > s] = P[b < (1-s)t^{-p}] \quad (29)$$

Therefore, by substituting for u in (23) the value $(1-s)t^{-p}$ and using (26) and (27),

$$P[Ts > t] = exp[-(t^p/\eta^p(1-s))^{\frac{\beta}{p}}] \quad (30)$$

or

$$P[Ts > t] = exp[-(1-s)^{-\frac{\beta}{p}}(\frac{t}{\eta})^{\beta}] \quad (31)$$

It is seen that (31) describes the reliability function of a Weibull random variable with: 1) the same shape factor β as the distribution of T; 2) A scale factor η_s expressed as follows as a function of s , the scale factor η of T and the exponent p:

$$\eta_s = \eta(1-s)^{\frac{1}{p}} \quad (32)$$

Thus, the problem stated in the beginning has also been solved in the Weibull distribution case.

Accordingly, the mathematical expectation of the first hitting time T_s is obtained:

$$E(T_s) = \eta(1-s)^{\frac{1}{p}}\Gamma(1+\frac{1}{\beta}) \quad (33)$$

Equation (33) can also be formulated as

$$E(T_s) = E(T)(1-s)^{\frac{1}{p}} \quad (34)$$

which is the same as in the Gamma-distribution case (18). Also, the expectation of the health index $HI(t)$ at time t can be derived from the expectation of the random coefficient b, assumed Fréchet distributed:

$$E(b) = \frac{1}{\eta^p}\Gamma(1-\frac{p}{\beta}) \quad (35)$$

Therefore

$$E(HI(t)) = 1 - E(b)t^p = 1 - (\frac{t}{\eta})^p\Gamma(1-\frac{p}{\beta}) \quad (36)$$

The quantiles of b can also be derived. The x-percent quantile is B_x:

$$B_x = \frac{1}{\eta^p(-lnx)^{\frac{p}{\beta}}} \quad (37)$$

In particular, the median (50-percent quantile) is given by:

$$B_{0.5} = \frac{1}{\eta^p(ln2)^{\frac{p}{\beta}}} \quad (38)$$

2.5. Estimation of Exponent p from Data

From (32), there follows, by taking logarithms,

$$\log(1-s) = p\log(\frac{\eta_s}{\eta}) \quad (39)$$

Therefore, after estimating η_s from the data sample for various values of s, the regression coefficient of $\log(1-s)$ with respect to $\log(\frac{\eta_s}{\eta})$ will provide an estimation of p. Also, taking (34) into account,

$$log(1-s) = p\log(\frac{E(T_s)}{E(T)}) \quad (40)$$

Therefore, in order to estimate p, it is equivalent to estimate $E(T_s)$ from the data samples corresponding to several values of s and then run the linear regression of $\log(1-s)$ with respect to $\log(\frac{E(T_s)}{E(T)})$. The regression coefficient (slope) is the best estimate of p. The same method applies in the Gamma distribution case since the dependence of E(T_s) on s is the same in both cases (see Section 2.3).

2.6. Incomplete Degradation Trajectories

Our method for obtaining an analytical form of the HI does not require run-to-failure condition monitoring data [1]. Let us

[1] If no failures are observed the HI has a different meaning as it is normalized with respect to the most degraded unit in the fleet.

consider u as the smallest threshold of $HI(t)$ observed for all units in the fleet. Then in equation (34), instead of considering the expected TTF, $E(T)$, we would consider the expected time to hit the common threshold $E(T_u)$. The revised equation would be:

$$E(T_s) = E(T_u)\frac{(1-s)^{1/p}}{(1-u)^{1/p}} \tag{41}$$

Where $E(T_s)$ is the sample arithmetic mean for each value $s > u$. When $u = 0$, this is equivalent to Eq. 34. The exponent p can be estimated from linear regression in

$$\log(1-u) - \log(1-s) = p(\log(E(T_u)) - \log(E(T_s))) \tag{42}$$

3. Case Studies

3.1. Turbofan

The New Commercial Modular Aero-Propulsion System Simulation (N-CMAPSS) dataset (Arias Chao, Kulkarni, Goebel, & Fink, 2021) offers comprehensive degradation trajectories of turbofan engines until failure. Among the dataset's eight subsets, we focus on DS003, characterized by a failure mode impacting the efficiency and flows of both low-pressure and high-pressure turbines.

The N-CMAPSS dataset characterizes degradation at the component level across initial, normal, and abnormal degradation stages. Consequently, an HI is calculated through a non-linear mapping of operational margins under reference conditions. System failure is determined when the HI reaches 0. The dataset also accounts for between-flight maintenance by allowing improvements in engine health parameters within specified limits. The ground truth HI is shown in Figure 2, and will be used to verify the findings of Section 2.3 and 2.4. Estimating the HI using condition monitoring data as highlighted in (Bajarunas et al., 2023) is also possible.

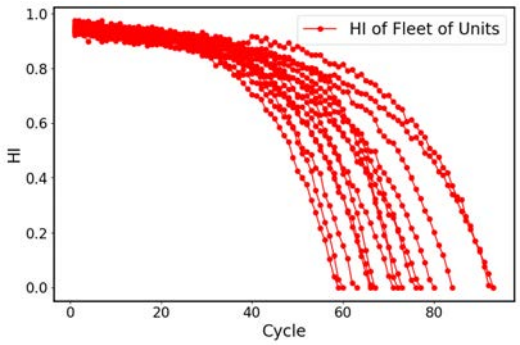


Figure 2. Observed HI in N-CMAPSS DS03 Dataset

The Akaike Information Criterion (AIC) (Akaike, 1974) was used to compare the goodness of fit with different probability distributions (Weibull, Gamma, Exponential), see Table 1. When a statistical model is used to represent the process that generated some data, some information is lost. The AIC, based on information theory, estimates the amount of information lost. It deals both with overfitting and underfitting by taking model simplicity into account as well as goodness of fit. The AIC is defined by

$$AIC = -2log(maxL) + 2P \tag{43}$$

where the term log(maxL) denotes the maximum value of the log-likelihood function, and P is the number of parameters in the model (for instance,for Weibull or Gamma, P is equal to 2). In our example, the best value of the AIC was obtained with the Gamma distribution for the time to failure as well as the time to reach level s for s ranging from 0 to 0.8. The AIC value for Weibull distribution is almost identical. In contrast, the AIC value for the exponential distribution is much higher.

Using the Maximum Likelihood Estimation technique, we estimated the best-fit Gamma parameters for various s thresholds. Figure 3 shows the estimated β_s and λ_s values for $s = [0, 0.1, 0.2, ..., 0.8]$. The results validate the conclusion presented in Section 2.3: the distribution of the first hitting time T_s shares the same shape factor $\beta = 52.83$ as the distribution of failure times T. Additionally, the rate parameter λ_s is a function of s and λ of T. We determined $p = 3.35$ following the description provided in Section 2.5. The wide confidence intervals of β_s and λ_s can be primarily attributed to the limited number of observations (15 run-to-failure curves), rather than to the choice of the Gamma distribution, which we have demonstrated to be the most suitable among the alternative distributions investigated.

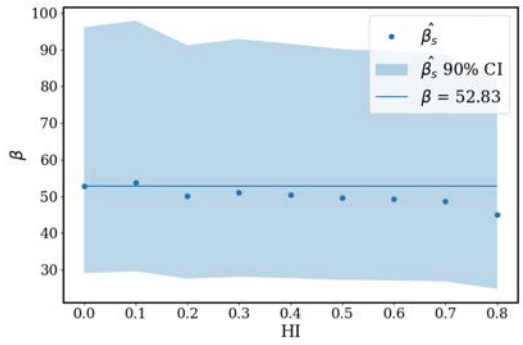
We then estimated the best-fit Weibull parameters for various s thresholds. In Figure 4, we estimated β_s and η_s using $s = [0, 0.1, 0.2, ..., 0.8]$. The results validate the conclusion presented in Section 2.4: the distribution of the first hitting time T_s shares the same shape factor $\beta = 7.32$ as the distribution of failure times T. Additionally, the scale parameter η_s is a function of s and η of T.

Figure 5 illustrates the mean, median, and 90% quantile of $HI(t)$, as described by equations (36) and (37). Notably, we observe that the median closely aligns with the ground truth HI within the dataset.

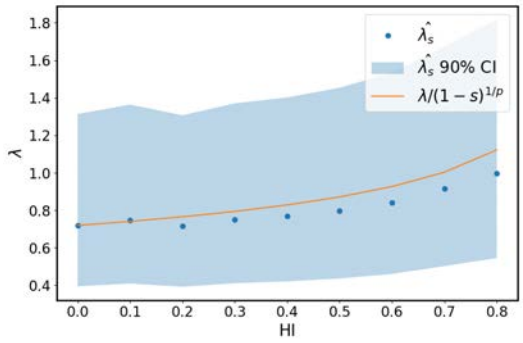
3.2. Battery

The methodology proposed in this study was further validated using a dataset obtained from the NASA Ames Prognostics Center of Excellence repository, specifically focusing on battery usage patterns (Bole, Kulkarni, & Daigle, 2014). This dataset includes information collected from individual 18650 LCO cells undergoing various charging and discharging cycles following randomized protocols.

Batteries commonly exhibit several physical aging mecha-



(a) β_s



(b) λ_s

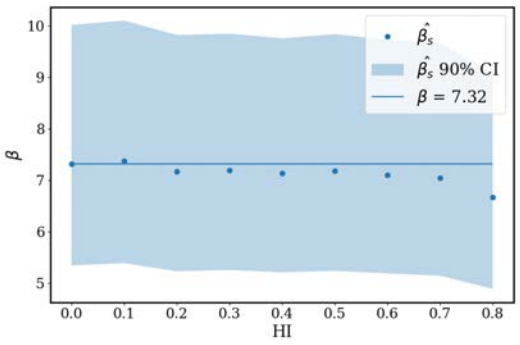
Figure 3. The Gamma distribution shape factor β_s and the rate parameter λ_s for various HI thresholds for N-CMAPSS dataset.

Table 1. AIC of distribution fits for CMAPSS turbofan case study.

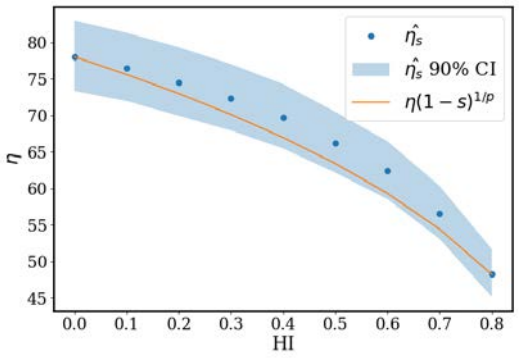
s	AIC Gamma	AIC Weibull	AIC Exponential
0	118	120	161
0.1	117	120	161
0.2	117	120	160
0.3	116	119	159
0.4	115	118	158
0.5	114	116	156
0.6	112	114	154
0.7	109	112	151
0.8	106	108	147
0.9	116	113	127

nisms such as graphite exfoliation, electrolyte loss, solid electrolyte interface layer formation, continuous thickening, and lithium plating, among others (Sui et al., 2021). These aging processes lead to two primary changes in battery behavior: capacity degradation and increased internal resistance. In this analysis, our focus will be on capacity degradation as the key health index for the batteries under investigation.

The HI of a battery is defined as the ratio between its current capacity and the nominal capacity (Q/Q_{nominal}). The battery's capacity can be determined by reference discharge cycles conducted at a constant current (I). The current capacity is calculated as the integral of current over the entire



(a) β_s



(b) η_s

Figure 4. The Weibull distribution shape factor β_s and the scale factor η_s for various HI thresholds for N-CMAPSS dataset.

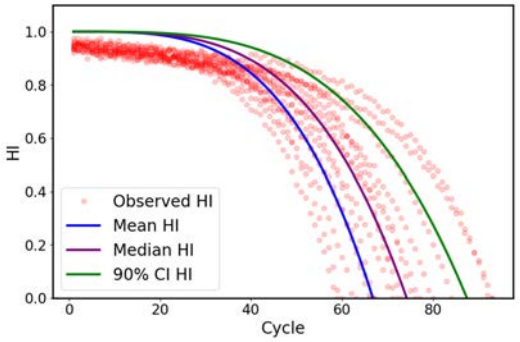


Figure 5. The mean, median, and 90% quantile of the health index obtained from Weibull distribution.

reference discharge cycle, denoted as $\int_t I$.

In this work, the failure of a battery (HI = 0) is defined once the capacity ratio is less than 60%. The initial HI of the battery is equal to the initial capacity ratio. Figure 6 shows the estimated HI of the NASA battery dataset.

The AIC values of three different distribution fits are shown in Table 2. The best fit was obtained with Gamma distribution for the time to failure as well as the time to reach level s for s ranging from 0 to 0.9. The AIC value for Weibull dis-

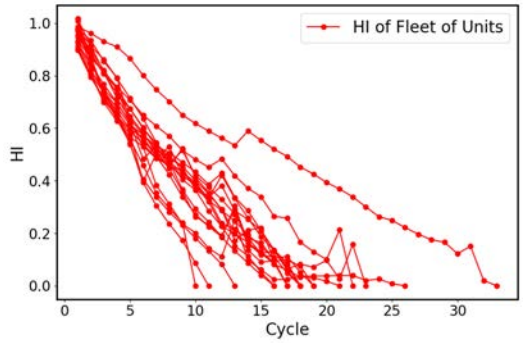
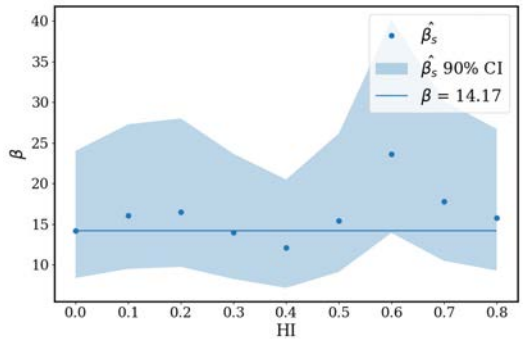


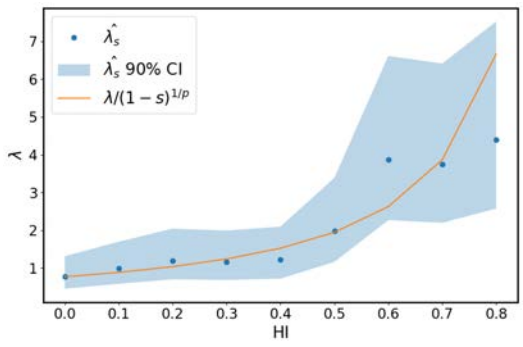
Figure 6. Observed HI in NASA battery dataset

tribution is almost identical,in contrast with the exponential distribution AIC, much higher.

We employed the Maximum Likelihood technique to estimate the best-fit Gamma parameters for various s thresholds. In Figure 7, we estimated β_s and λ_s using $s = [0, 0.1, 0.2, ..., 0.8]$. Once more, we illustrate that a reasonably good approximation for the shape parameter β_s of the first hitting time is the shape parameter β of T. Following the estimation of $p = 0.94$, we demonstrate that the rate parameter λ_s varies with s and λ. Since $p < 1$, the HI curve is now convex, as observed.



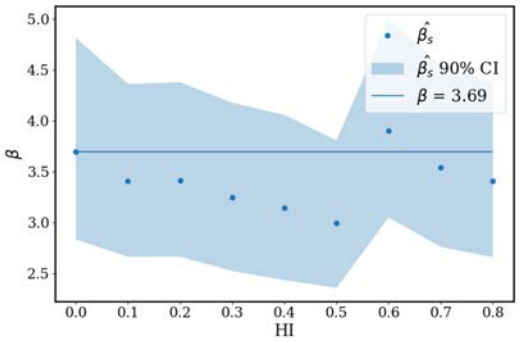
(a) β_s



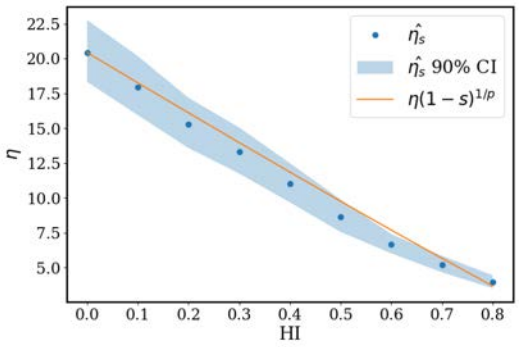
(b) λ_s

Figure 7. The Gamma distribution shape factor β_s and the rate parameter λ_s for various HI thresholds for NASA battery dataset.

The best-fit Weibull parameters for various s thresholds are shown in Figure 8. Once again, we show that a reasonably good approximation for the shape parameter of the first hitting time β_s is the shape parameter β of the failure time T and that the scale parameter η_s varies with s and η as expected.



(a) β_s



(b) η_s

Figure 8. The Weibull distribution shape factor β_s and the scale factor η_s for various HI thresholds for NASA battery dataset.

Table 2. AIC of other distribution fits for NASA battery case study.

s	AIC Gamma	AIC Weibull	AIC Exponential
0.0	118	121	151
0.1	111	118	146
0.2	104	112	140
0.3	102	108	135
0.4	97	102	128
0.5	84	93	118
0.6	67	75	109
0.7	62	70	99
0.8	54	60	89
0.9	44	50	71

4. Sensitivity Analysis

Sensitivity analysis has been conducted on the N-CMAPSS dataset, to investigate the effect of the exponent p in the parametric model of the health index.

For the Gamma case, it is immediate from (16) that, for given

s, λ_s is a decreasing function of p (for p greater than, or equal to 1). In the limit of p going to infinity, λ_s converges to λ. For the Weibull case, a similar conclusion is drawn, but instead from (32) it follows that, for given s, η_s is an increasing function of p.

From (18) and (34) it follows that for both considered distributions the average time to reach threshold s, $E(T_s)$, is an increasing function of p, as illustrated in Figure 9.

For both distributions, when p increases, the average value of the HI is first higher than, and subsequently (for greater values of the time variable t), lower than, the HI corresponding to a lower value of p. Increasing p corresponds to delaying the decrease in HI, i.e., delaying the onset of the degradation; but, once the degradation occurs, it is more sudden. See Figure 10 for an illustration.

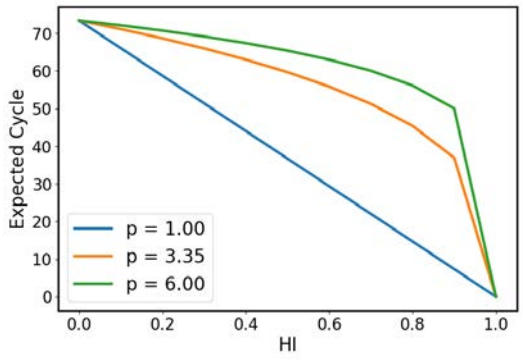
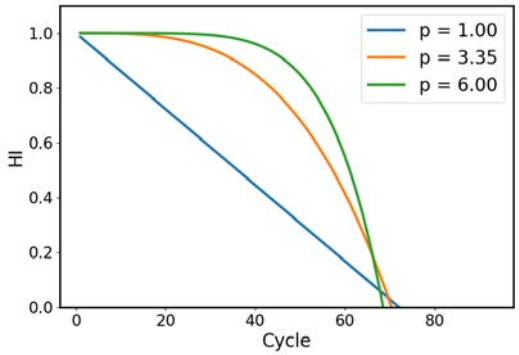


Figure 9. Gamma and Weibull distribution $E[T_s]$ as a function of s for three values of p. N-CMAPSS dataset.

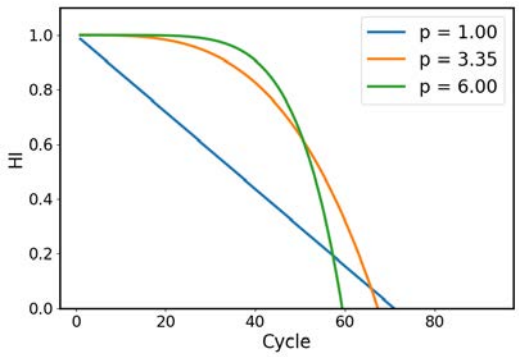
5. Conclusion and Perspectives

This study has successfully addressed the problem of analytically modeling health indices (HI) in cases where the time-to-failure follows either a Gamma or Weibull distribution. By leveraging observed health index trajectories over time and specifically the failure times, we have derived an analytical form for the health index that is consistent with these observations. Additionally, we provided an analytical expression for the distribution of the time to reach any intermediate degradation level.

The availability of closed-form expressions for the health index is highly beneficial for implementing predictive maintenance strategies, particularly for estimating the remaining useful life (RUL) distribution. Furthermore, once a health index function is derived for a particular application, it can potentially serve as a foundation for similar applications, such as the same asset under different operating conditions or a slightly modified asset. Without an analytical characterization, a new health index would need to be learned from scratch for each new dataset.



(a) Gamma distribution $E[HI(t)]$ as a function of t for three values of p. N-CMAPSS dataset.



(b) Weibull distribution $E[HI(t)]$ as a function of t for three values of p.

Figure 10. Sensitivity to various p for the turbofan case study. N-CMAPSS dataset.

Future work could extend this approach to other TTF distributions and other HI formulations by applying the general methodology outlined in Section 2.2. Additionally, an important extension of this work could be the use of quantile regression and extrapolation of the HI from individual degradation trajectories. More broadly, the analytical health index approach represents a significant advancement in survival analysis, offering opportunities to integrate machine learning techniques, particularly 'deep survival' methods, with traditional reliability engineering.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5.

Atamuradov, V., Medjaher, K., Camci, F., Zerhouni, N., Dersin, P., & Lamoureux, B. (2020). Machine health indicator construction framework for failure diagnostics

and prognostics. *Journal of signal processing systems*, *92*, 591–609.

Bajarunas, K., Baptista, M., Goebel, K., & Chao, M. A. (2023). Unsupervised physics-informed health indicator estimation for complex systems. In *Annual conference of the phm society* (Vol. 15).

Bole, B., Kulkarni, C. S., & Daigle, M. (2014). Adaptation of an electrochemistry-based li-ion battery model to account for deterioration observed under randomized use. In *Annual conference of the phm society* (Vol. 6).

Dersin, P. (2023). *Modeling remaining useful life dynamics in reliability engineering*. CRC Press,Taylor and Francis.

Djeziri, M. A., Benmoussa, S., & Zio, E. (2020). Review on health indices extraction and trend modeling for remaining useful life estimation. *Artificial Intelligence Techniques for a Scalable Energy Transition: Advanced Methods, Digital Technologies, Decision Support Tools, and Applications*, 183–223.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. de la Soc. Polonaise de Math.*.

Hsu, C.-C., Frusque, G., & Fink, O. (2023). A comparison of residual-based methods on fault detection [Conference Paper]. In C. S. Kulkarni & I. Roychoudhury (Eds.), *Proceedings of the annual conference of the phm society 2023* (Vol. 15). s.l.: PHM Society. (15th Annual Conference of the Prognostics and Health Management Society (PHM 2023); Conference Location: Salt Lake City, UT, USA; Conference Date: October 28 - November 2, 2023) doi: 10.3929/ethz-b-000636893

Llera, A., & Beckmann, C. F. (2016). Estimating an inverse gamma distribution. *arXiv:1605.01019v2*.

Nachlas, J. (2017). *Reliability engineering- probabilistic models and maintenance methods, 2d edition*. CRC Press,Taylor and Francis.

Qin, Y., Yang, J., Zhou, J., Pu, H., & Mao, Y. (2023). A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery rul prediction. *Advanced Engineering Informatics*, *56*, 101973. doi: https://doi.org/10.1016/j.aei.2023.101973

Ramos, P., Louzada, F., Ramos, E., & Dey, S. (2020). The fréchet distribution: Estimation and application-an overview. *Journal of Statistics and Management Systems*, *23*(3), 549-578.

Roman, D., Saxena, S., Robu, V., Pecht, M., & Flynn, D. (2021). Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence*, *3*(5), 447–456.

Sui, X., He, S., Vilsen, S. B., Meng, J., Teodorescu, R., & Stroe, D.-I. (2021). A review of non-probabilistic machine learning-based state of health estimation techniques for lithium-ion battery. *Applied Energy*, *300*, 117346.

Yang, F., Habibullah, M. S., & Shen, Y. (2021). Remaining useful life prediction of induction motors using nonlinear degradation of health index. *Mechanical Systems and Signal Processing*, *148*, 107183.

Ye, Z., & Yu, J. (2021). Health condition monitoring of machines based on long short-term memory convolutional autoencoder. *Applied Soft Computing*, *107*, 107379.

BIOGRAPHIES

Dr.Pierre Dersin: Dr. Dersin is currently Adjunct Professor at Luleå University of Technology (Sweden) in the Operation and Maintenance Engineering Division. He is also the president and founder of Eumetry sas, Louveciennes,France, a consulting firm in the fields of RAMS, PHM and AI. He holds a Ph.D. in Electrical Engineering from the Massachusetts Institute of Technology (MIT),as well as a Master's degree in Operations Research, also from MIT, and math and E.E. degrees from Université Libre de Bruxelles (Belgium). From 1990 to 2021, he was with Alstom Transport, France, where he occupied several technical and managerial positions, including RAM (Reliability-Availability-Maintainability) Director and RAM Master Expert,and founded the "RAM Center of Excellence".In 2015, he contributed to the launch of the predictive maintenance activity at ALSTOM and became PHM (Prognostics and Health Management) Director of ALSTOM Digital Mobility, St-Ouen, France. From 2014 to 2018, he was also the co-director of the joint Alstom-Inria Research Lab on Digital Mobility, and supervised several Ph.D. theses. Prior to joining Alstom, he was employed in the USA ,first as a research scientist in the US DOE large-scale system effectiveness program; and subsequently,with Belgian engineering firm Fabricom's US subsidiary, involved with fault detection and diagnostics in industrial systems. He has contributed a number of conference and journal papers in the fields of RAMS, PHM, automatic control, electric power systems, and AI. He was the keynote speaker at the 2014 European Conference of the PHM Society . Dr.Dersin is the author of the book " Modeling Remaining Useful Life Dynamics in Reliability Engineering" , CRC Press, Taylor and Francis, June 2023. His current interests focus on the confluence between RAMS and PHM as well as complex systems resilience and asset management.

Kristupas Bajarunas: Kristupas Bajarunas is a Ph.D. candidate at Delft University of Technology (Netherlands) and Zurich University of Applied Sciences (Switzerland). He holds an M.Sc. in Machine Learning from the Royal Institute of Technology (Stockholm). His current research is concerned with developing generic hybrid models for prognostics.

Dr. Manuel Arias Chao: Dr. Arias Chao is a Senior Lecturer at Zurich University of Applied Sciences and a visiting researcher at the Air Transport and Operations of the Delft University of Technology in the Netherlands. He has a PhD in Physics-informed Machine Learning for Prognostics and Health Management from ETH Zurich, a Master's degree in

Thermal Power from Cranfield University, and a Bachelor's degree in Aeronautical Engineering from the Technical University of Madrid. Manuel has gained valuable industrial and research experience as a visiting researcher at the Diagnostics & Prognostics Group at NASA Ames, Thermodynamics & Performance Lead Engineer at General Electric and ALSTOM Power, and Aero Engine Maintenance Engineer at ITP. In his current role, Manuel also co-leads the Expert Group Smart Maintenance from the Swiss Alliance for Data-Intensive Services. He focuses on teaching & research for a broad range of application fields, including power generation, marine and aircraft propulsion, and manufacturing equipment.

Anomaly Detection of a Cooling Water Pump of a Power Plant Based on its Virtual Digital Twin Constructed with Deep Learning Techniques

Miguel A. Sanz-Bobi[1], Sarah Orbach[1], F. Javier Bellido-López[1], Antonio Muñoz[1], Daniel Gonzalez-Calvo[2], and Tomas Alvarez-Tejedor[2]

[1]*Institute for Research in Technology, ICAI School of Engineering, Universidad Pontificia Comillas,*

Santa Cruz de Marcenado 26, 28015 Madrid, Spain
masanz@comillas.edu

[2] *Enel Green Power and Thermal Generation, Endesa - Gas Maintenance Iberia, Ribera del Loira 60, 28042 Madrid, Spain*
daniel.gonzalezc@enel.com

ABSTRACT

This paper aims to explore the use of recent approaches of deep learning techniques for anomaly detection of potential failure modes in a cooling water pump working in a gas-combined cycle in a power plant. Two different deep learning techniques have been tested: neural networks and reinforcement learning. Two virtual digital twins were developed with each family of deep learning techniques, able to simulate the behavior of the cooling water pump in the absence of pump failure modes. Each virtual digital twin consists of several models for predicting the expected evolution of significant behavior variables when no anomalies exist. Examples of these variables are bearing temperatures or vibrations in different pump locations. All the data used comes from the SCADA system. The main features and hyperparameters in the virtual digital twins are presented, and demonstration examples are included.

1. INTRODUCTION

The early anomaly detection of failure modes in a power plant is a key factor in mitigating their effects on its operation, maintenance, and, in general, potential costs not planned. This problem has been studied intensively in the scientific literature for some time. Today, the availability of a large amount of data and the development of different machine learning techniques have propitiated their increasing use for early anomaly detection. References (Chavan & Yalagi, 2023), (Pang, Shen, Cao, Van Den Hengel (2021)), or (Nassif, Talib, Nasir & Dakalbab (2021)) are some examples of literature reviews in this area.

Also, anomaly detection with respect to the expected normal behavior is a crucial input for a data-driven, efficient prognostics and health management program (PHM) as is inferred from references (Maior, Araújo, Lins, Moura &. Droguett, 2023), (Ochella, Shafiee & Dinmohammadi, 2021) and (Calvo-Bascones, Sanz-Bobi & Welte, 2021). These references are based on machine learning techniques as the main tools to reach their objectives.

In line with these principles, this paper presents a digital twin of a cooling water pump (CWP) working in a power plant able to emulate normal behavior through a set of characteristic variables when the pump is in normal operation. The variables predicted are those that were considered important for the detection of anomalies that could cause failure modes. The digital twin was developed using two different families of deep learning techniques (Bishop & Bishop, 2024): neural networks and reinforcement learning. The use of reinforcement learning techniques in this field is less known, and this paper explores its potential by comparing the results obtained with a more extended method based on deep neural networks.

The paper is organized as follows: Section 2 describes the foundations of the study. Section 3 presents the digital twins developed and the methodology used. Section 4 shows the anomaly detection results based on the digital twins. Finally, Section 5 presents the more relevant conclusions reached.

2. STUDY FOUNDATIONS

The objective of the analysis presented in this paper is to detect anomalies as soon as possible in the behavior expected for a Cooling Water Pump (CWO) working in a combined

cycle of a power plant. Even when the paper is focused on the case of a CWP, the idea is to develop a procedure that can be easily extended to other components in the power plant. A CWP (Bowman & Bowman, 2021) is an important component whose objective is to cool the steam from the water turbine in an enthalpic process that contributes to the improvement of the water-steam cycle efficiency in the power plant. The process of monitoring if the pump is working as expected in normal behavior for any working condition is based on the consideration of the most typical failure modes that could appear in this type of pump. A Failure Mode and Effects Analysis (FMEA) (Huang, You, Liu & Song, 2020) suggested the main observable variables that could indicate the presence of an anomaly and that can be summarized by monitoring the vibrations in the pump axes, temperature in bearings, and currents in the electrical motor. As it is known, all of these potential failure modes have an important critical impact on keeping the pump in healthy condition.

This information has been used to support the development of a virtual digital twin to predict, in any working condition, the expected values of the vibrations in the axes of the pump, the temperature in its main bearings, and the currents in the electrical motor. This virtual digital twin is based on several models that characterize the relationships between variables to monitor for possible anomalies and the working conditions of the power plant. The list of variables predicted by the models developed follows:

- Prediction of vibration in axis X
- Prediction of vibration in axis Y
- Temperature of the bearing on the electrical motor side
- Temperature of the bearing on the pump side
- Temperature of pump thrust bearing
- Current in the electrical motor

The inputs to all the models correspond with the power generated by the steam turbine of the combined cycle power plant, which is the most important flow to cool, and the temperature in the electrical motor representing the work developed by the pump. The CWP is a high-pressure pump that is horizontal, centrifugal, and multistage. The pump and its motor are mounted on a common structural steel bedplate. Its behavior is monitored from a control room of the power plant where the variables measured in the CWP are accessible.

3. DIGITAL TWIN MODELS

Two redundant virtual digital twins (Jones, Snider, Nassehi, Yon & Hicks, 2020) were developed for the CWP. Both aim to simulate the CWP when it works in normal conditions. The emulation of the expected values for anomaly detection of the target variables is based on a double redundant strategy that uses two different families of deep learning algorithms supporting the models cited: deep learning neural networks and deep reinforcement learning. The datasets used for the creation of the models behind the digital twin correspond to three years of the CWP operation that here will be called year 1, year 2, and year 3. Year 1 will be used for learning the relationships to model, and the other two years will be used for checking the behavior of the digital twin of the pump, simulating its behavior. Python is the programming language used in both versions of the CWP digital twin.

The following subsections will present the results reached by these two types of algorithms.

3.1. CWP digital twin based on Deep Learning Neural Networks (DLNN)

As previously mentioned, six models were created using deep learning neural networks. The procedure followed was similar in all the cases; for this reason, only one case will be described as an example of the method followed. If the whole set of models is used, more than one type of anomaly related to one failure mode could be detected. The example case described here is the estimation of the bearing temperature at the mechanical axis on the side of the connection to the electrical motor. The dataset used for training is Year 1. The input variables were the power generated by the steam water and the temperature in the electrical motor, which represents the working conditions of the CWP. After preprocessing and scaling the data, the *Optuna* open software tool (Akiba, Sano, Yanase, Ohta & Koyama, 2019) was used to find a convenient architecture and hyperparameters of the deep learning neural network. Tables 1 and 2 present the main characteristics of the neural network architecture and the most significant hyperparameters.

Table1. Deep Learning Neural Network. Architecture

Layer	Layer type	Number of neurons	Activation function
1	Dense	40	Sigmoid
2	Dense	25	Sigmoid
4	Dense	1	Sigmoid
4	Output	1	Linear

Table 2. Algorithm Implemented

Language	Python
Main Library	Keras and Tensorflow
Loss function	Mean Squared Error
Optimiser	Adam, learning_rate=0.001, epsilon=1^{e-8}
Training	Epochs=500, steps per epoch=100

Figure 1 shows the result of the relationship learned between the input and output variables based on data from Year 1. Both real and predicted values of the bearing temperature are very close. Their difference or error is in Figure 2. It shows that the most part of the error is the interval [-1, 1] ºC. It suggests a good simulation performance for this part of the digital twin. The axis X is in samples separated by 10 minutes.

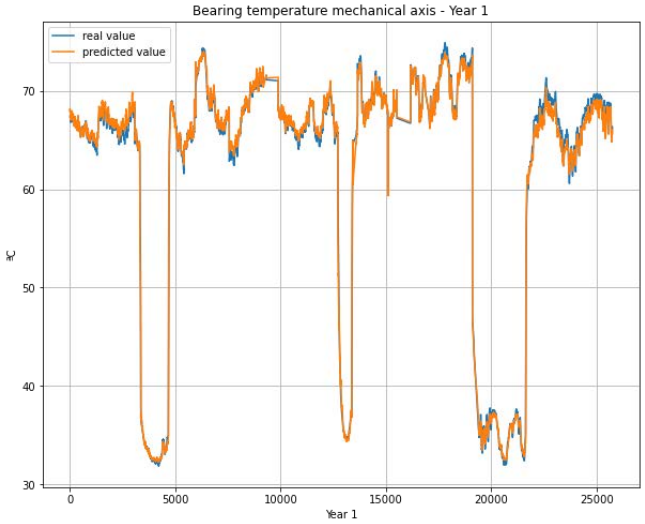


Figure 1. Real and predicted values for the bearing temperature using Year 1 data.

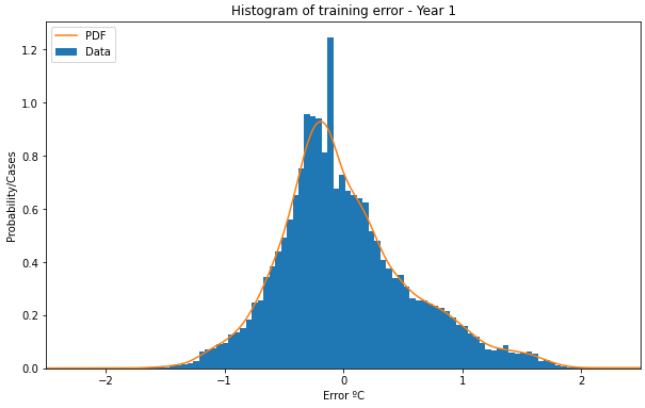


Figure 2. Error observed between real and predicted values for the bearing temperature using Year 1 data.

As a confirmation of the goodness of the model obtained to predict the bearing temperature, Figure 3 presents how good the prediction of this variable is when data used were not included during the creation of the model. Once again, the real and predicted values are very close, concluding that the model created is valid to simulate the bearing temperature in the normal behavior of the CWP. Also, this confirms that no overfitting issues are present. The errors observed are in the same range of values observed with the training dataset, and the same conclusion is reached for the Year 2 dataset. This confirms that this model can be used as a virtual twin of part of the CWP.

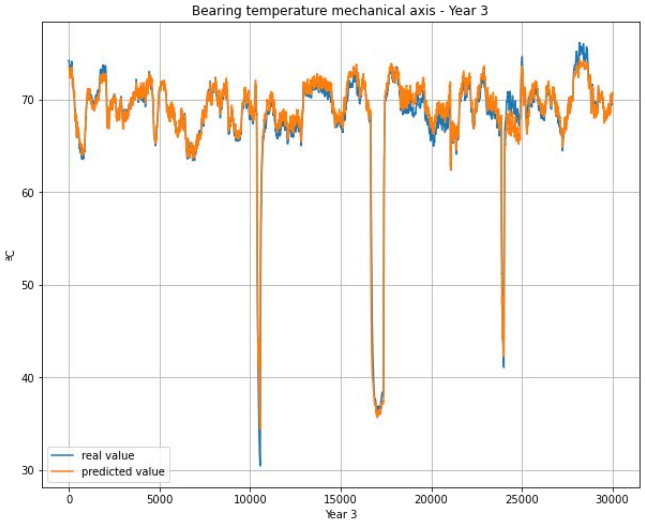


Figure 3. Real and predicted values for the bearing temperature of the dataset Year 3 not used in training.

3.2. CWP digital twin based on Deep Reinforcement Learning (DRL)

Once the CWP digital twin was developed using DLNN, a completely different type of algorithm was studied to cover the same objective. The idea was to explore Reinforcement Learning (RL) techniques for elaborating a digital twin of the CWP. At present, these techniques are not used very often in diagnosing industrial processes, and the number of publications about them is very limited.

Reinforcement Learning (Sutton & Barto, 2018) is a technique where an agent acts in an environment. It has a state, and it can make an action. After each action, the environment provides it with its new state and a reward corresponding to how good the action was. Therefore, the agent learns the parameters of a Quality function and makes new actions according to it. The agent has a multidimensional state space and a multidimensional action space. Figure 4 represents a schema of the basic cyclic process used in RL.

Figure 4 represents the correspondence between the main elements of RL. The objective is to build the same models described for the case DLNN. Here, the state is the input data used by the models. The action is the prediction of the behavior of the pump, considering its different working conditions. The environment gives a reward to penalize how far the prediction is from the real value and gives a new state, which is the new entrance data. The action space is continuous because a real value represents it. There are several RL algorithms; however, due to the continuous nature

of the problem to be managed, a Twin Delayed Deep Deterministic Policy Gradient algorithm (TD3) was selected. Details of the method and a pseudocode can be found in (Fujimoto, van Hoof & Meger, 2018).

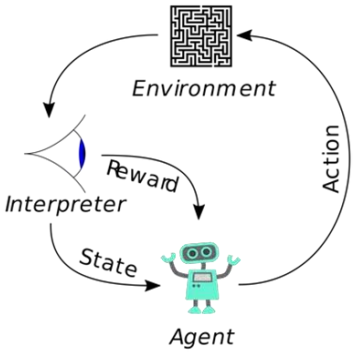


Figure 4. Basic learning cycle of Reinforcement Learning.

With the TD3 algorithm, the agent needs two deep neural networks to decide on its action. The PolicyNetwork determines an action to take. It takes the input data and returns the action, and the Q_network determines the Q_value of the action. It takes the input data and the action and returns the Q_value. During the training phase of the algorithm, the gradient for training the Q_network is calculated based on a linear combination between the reward and the prediction of the algorithm. The PolicyNetwork is trained based on the variation of the Q_network. The Q_Network weights are initialized between -3 and 3. The PolicyNetwork weights are initialized between -0,3 and 0,3. The networks are composed of three linear dense layers with a *relu* activation function.

Table 3. TD3 Hyperparameters

Algorithm Hyperparameter	Value
Training episodes	100
Steps per episode	100
Exploration factor	0.1
Replay buffer size	1032
Batch size	1024
Delayed steps for updating the policy network and target networks	10
Size of hidden layers for networks	64
Learning rate Q_network	3e-4
Learning rate Polocy_network	3e-4
Reward scale	100.

Table 3 presents the values used for the main hyperparameters of the TD3 algorithm. *Keras* and *tensorflow* were used for the implementation of the TD3 algorithm. The reward in RL is essential to guide the correct learning process. The reward design is based on a function of 6 levels depending on the absolute difference between the real and predicted values observed for the output variable of the model. The reward ranges from 100 for differences lower than 0.001, till -1500 for differences higher than 0.3.

Figure 5.a shows the result of the relationship learned between the input and output variables based on data from Year 1. Both real and predicted values of the bearing temperature are very close. Their difference or error is in Figure 5.b It shows that the most part of the error is the interval [-1, 1] ºC. It suggests a good simulation performance for this part of the digital twin.

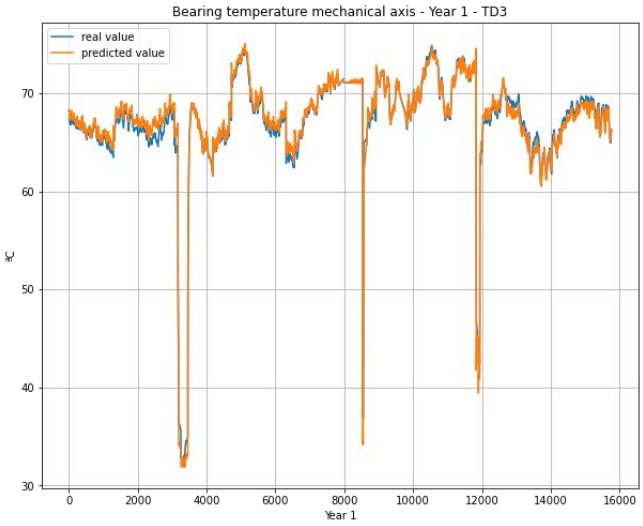


Figure 5.a. Real and predicted values for the bearing temperature using Year 1 data and TD3 algorithm.

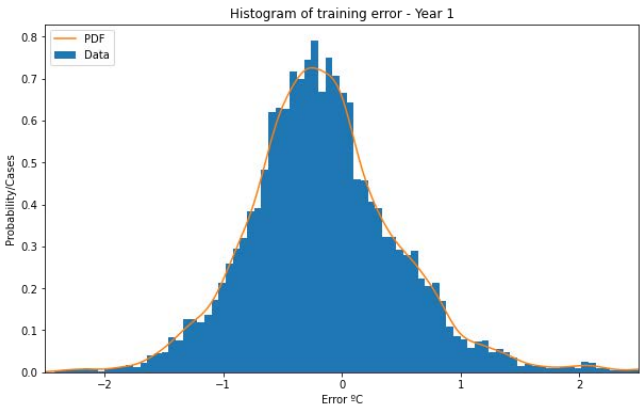


Figure 5.b. Error observed between real and predicted values for the bearing temperature using Year 1 data and TD3 algorithm.

Figure 6 confirms the model goodness using RL to predict the bearing temperature when the data used, Year 2, were not included in the model training process. Once again, the real and predicted values are very close, concluding that the model created with the RL algorithm TD3 is also appropriated to simulate the bearing temperature in the normal behavior of the CWP. The errors observed are in the same range of values observed with the training dataset, and the same conclusion is reached for the Year 3 dataset. This confirms that this model can be used as a virtual twin of part of the CWP.

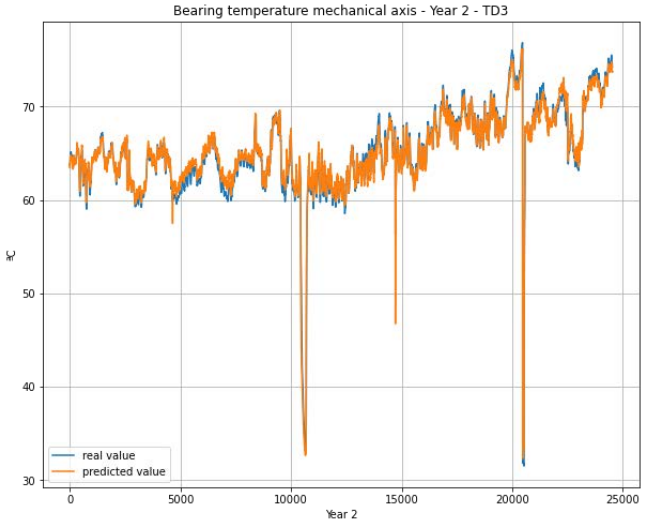


Figure 6. Real and predicted values for the bearing temperature using Year 2 data and TD3 algorithm.

3.3. Comparative results of both CWP digital twins: DLNN and DRL

Once both virtual digital twins were obtained to simulate the CWP performance, one of the objectives of this study was reached, which was the comparison between the use of DLNN and DRL techniques. As mentioned, the use of DRL for this type of problem is not too extended. The results obtained have demonstrated that DRL is a reasonable option in terms of simulation of the behavior of a real industrial component. Table 4 shows the mean and standard errors obtained in °C degrees in all the cases studied with both deep learning methods. It can be observed that the values are within the accuracy of any temperature sensor used in industry, and both methods can be used with similar confidence for detecting deviations concerning the normal behavior expected.

However, the main objective of this study is the early detection of possible anomalies that can cause failures. The next section will show how the digital twins can be used for that.

Table 4. Mean and standard errors obtained

	Year 1	Year 2	Year 3
Mean_DLNN	0,03600	-0,26709	0,16897
Std_DLNN	0,71556	1,36557	1,24222
Mean_TD3	-0,14583	-0,04837	-0,04837
Std_TD3	0,68032	0,04842	0,99234

4. ANOMALY DETECTION AND RISK ASSESSMENT

The digital twins described in the previous section and their good performance permit the application of a redundant strategy for robust anomaly detection. It is important to note that the algorithms used for both digital twins are completely different, even when they observe the same information. It seems clear that if both coincide in detecting an anomaly in the behavior expected, its certainty should be high. Also, if both observe normal behavior. In the case of a discrepancy, careful monitoring must be observed for the new coming data. Redundancy is key for preventing false alarms in anomalies detected.

Another important point to note is that the digital twins were developed to learn the normal behavior in the CWP operation expressed by several values of variables observed in the SCADA system. If, for some of these variables, the value observed is not similar to the value predicted by the digital twin, then an abnormal behavior is present that has to be investigated. In fact, the variables observed and predicted were selected as direct indicators of the presence of possible failure modes.

The models obtained by DLNN and DRL techniques simulate very well the normal behavior expected for the output variables; however, they show small discrepancies between real and predicted values, such as those presented in Table 4. In order to prevent false alarms in both cases, confidence bands, according to the error observed in the training models, were defined for monitoring new data differently from those used for training. These confidence bands were adopted around the ±3 times the standard deviation of the error observed in training. Any new prediction inside these bands means that there is no behavior different from this expected for the variable predicted.

Figure 7 shows an example of the upper and lower confidence bands (straight lines) for the error observed in the prediction of Year 3 data that were not used for learning. In this figure, the error is inside the confidence bands, concluding that the temperature in this pump bearing is as expected and no anomaly is present. The same approach of confidence bands is applied to any model inside each digital twin.

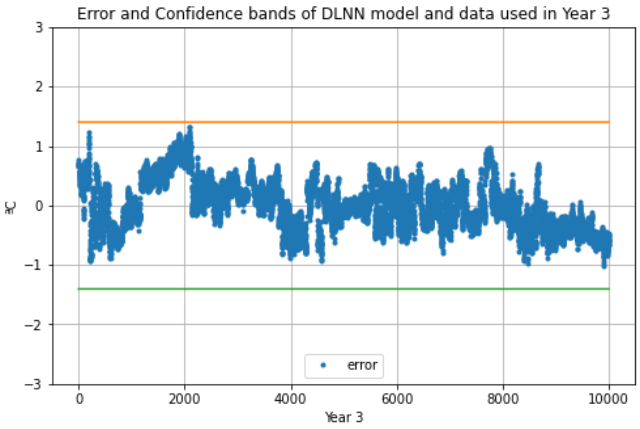


Figure 7. Error inside confidence bands: the behavior observed is similar to the expected one.

The main objective of this section is to present the ability of both digital twins to detect anomalies. In order to reach this goal, an artificial abnormal behavior was introduced in Year 3, simulating an increase in the temperature in the bearing studied in the previous section that could be the result of an incipient failure mode due to the wearing of balls in the bearing or weak lubrication. In this case, only an isolated possible failure mode is considered, leaving the problem of simultaneous failure modes open for further studies. Figure 8 shows the predicted and real values of the bearing temperature in the mechanical axis of the CWP digital twin based on DLNN. In the right part of the figure, there is a significant deviation between real and predicted values that alert about abnormal temperature behavior for the observed working conditions.

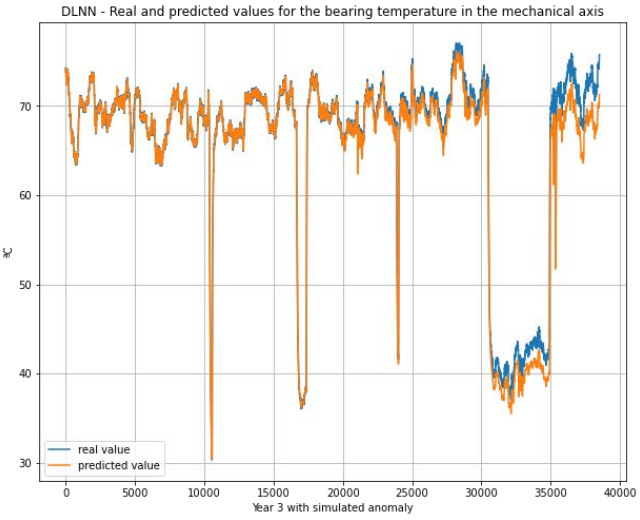


Figure 8. Real and predicted values by the DLNN digital twin. An anomaly is detected at the end of the period.

The deviation with respect to the normal behavior is observed in detail in Figure 9, where the confidence bands of the error are also represented. On the left part of the figure, there is a very clear deviation with respect to the normal behavior expected, meaning that the temperature in the bearing is higher than normal for the current working conditions. Also, it is possible to observe the current fingerprint of the detected anomaly, keeping only the information in Figure 9 that is out of the upper confidence band. This is presented in Figure 10, where an increase of about 2.5 ºC degrees in the last 50 days (6000 samples) is observed, and its trends will be called "risk of the failure mode" in this paper. This information is precious for an approach to implement a data-driven maintenance program.

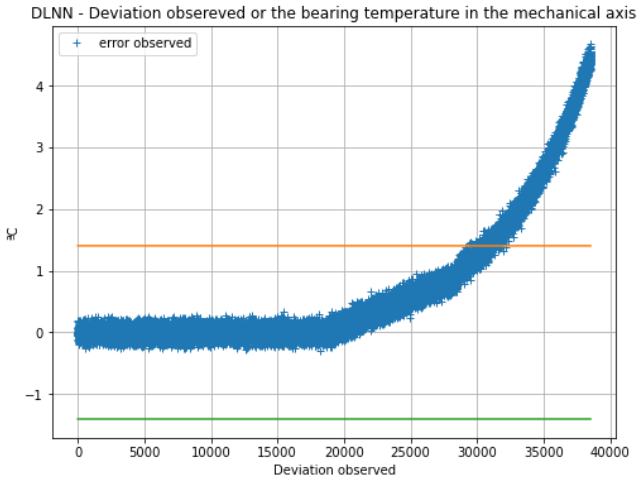


Figure 9. Error inside confidence bands: the behavior observed is similar to the expected one. DLNN case.

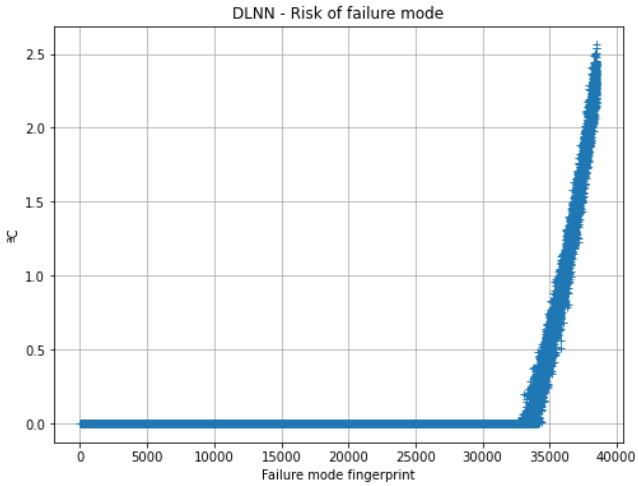


Figure 10. Risk of failure mode based on DLNN digital twin.

The performance of the digital twin based on DRL is similar to that described for the digital twin based on DLNN. Figures 11, 12, and 13 present these results.

Figure 14 presents the superposition of the risk presented in Figures 10 and 13. The objective is to check if some digital twin detects the anomaly condition presented earlier. According to this figure, both virtual digital twins are able to detect the anomaly at the same time. The error observed from the DLNN digital twin seems to be slightly higher, but in any case, it is not significant in ºC units. This confirms the reliability and robustness of the method for anomaly detection based on these digital twins built using different deep learning techniques.

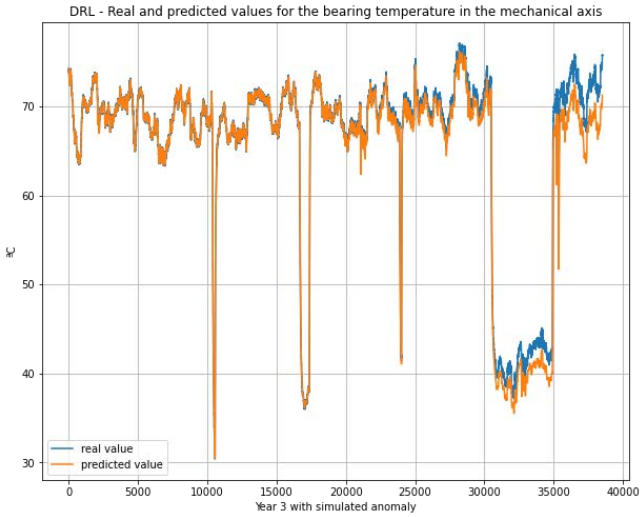


Figure 11. Real and predicted values by the DRL digital twin. An anomaly is detected at the end of the period.

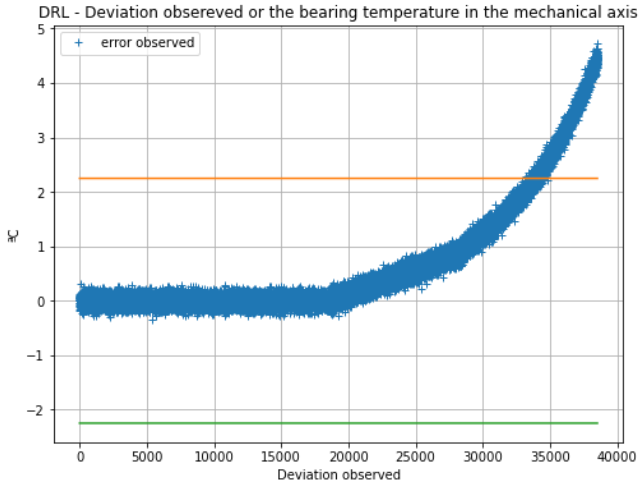


Figure 12. Error inside confidence bands: the behavior observed is similar to the expected one. DRL case.

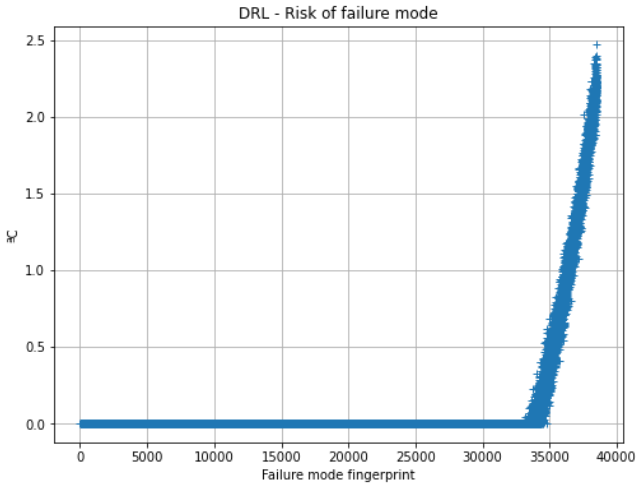


Figure 13. Risk of failure mode based on DRL digital twin.

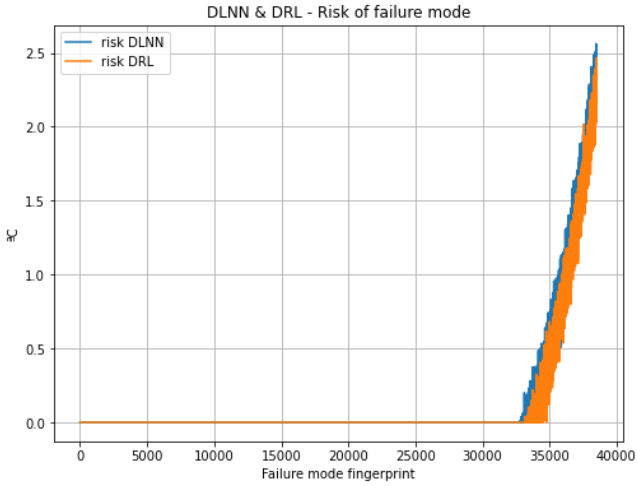


Figure 14. Comparison of risks observed with DLN and DRL digital twins.

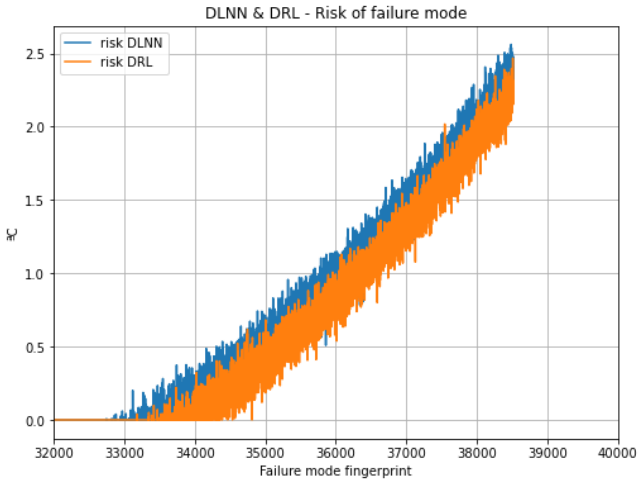


Figure 15. Zoom of the interest zone of Figure 14.

5. CONCLUSION

This paper has presented two virtual digital twins for anomaly detection in a CWP. The objective of creating two digital twins was double. First, the performances between DLNN and DRL were compared because the use of DRL is not well-known yet in this field, and it could have some advantages over the well-known DLNN method because it would need less data for training. The conclusion is that DRL techniques can be used as an alternative option for the DLNN. Second, using two digital twins based on different techniques could robust the anomaly detection process, preventing false alarms. This was verified and confirmed with an example of isolated failure. Additionally, the fingerprint of the detected anomaly can be used as an indicator of risk for a failure mode and alert maintenance people about this fact, giving the basis for a data-driven approach supporting the maintenance and asset management of industrial processes.

The results of this paper open several future studies, such as the analysis of the performance of the digital twins when several failure modes appear simultaneously and the propagation of their effects. Also, the use of the profiles of the risk of failure modes and their integration in maintenance promises to implement new maintenance plans.

ACKNOWLEDGEMENT

The study has been developed with the scientific and economic support of the ENDESA Chair of Artificial Intelligence Applications to Data-driven Maintenance.

REFERENCES

Akiba T., Sano S., Yanase T., Ohta T. & Koyama M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Pages 2623–2631. doi:10.1145/3292500.3330701

Bishop C. & Bishop H. (2023). *Deep Learning - Foundations and Concepts*. Springer Cham.doi:10.1007/978-3-031-45468-4

Bowman, C.F., & Bowman, S.N. (2021*). Engineering of Power Plant and Industrial Cooling Water Systems*. CRC Press. doi: 10.1201/9781003172437

Calvo-Bascones P., Sanz-Bobi M.A. & Welte T.M. (2021). Anomaly detection method based on the deep knowledge behind behavior patterns in industrial components. Application to a hydropower plant. *Computers in Industry*, Vol. 125, 103376. doi: 10.1016/j.compind.2020.103376.

Chavan, V.D. & Yalagi, P.S. (2023). A Review of Machine Learning Tools and Techniques for Anomaly Detection. In: *Choudrie, J., Mahalle, P.N., Perumal, T., Joshi, A. (eds) ICT for Intelligent Systems. ICTIS 2023. Smart Innovation, Systems and Technologies*, Vol 361. Springer.

Fujimoto S., van Hoof H. & Meger D (2018). Addressing function approximation error in actor-critic methods. Proceedings of the International Conference on Machine Learning. Vol. 80 Proceedings of the 35th International Conference on Machine Learning, PMLR 80: 1587-1596.

Huang J., You J., Liu H. & Song M (2020). Failure mode and effect analysis improvement: A systematic literature review and future research agenda. Reliability Engineering & System Safety.Vol. 199,106885. doi:10.1016/j.ress.2020.106885

Jones D., Snider C., Nassehi A., Yon J. & Hicks B. (2020) Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, Vol. 29, Part A, pp 36-52. doi:10.1016/j.cirpj.2020.02.002.

Maior C. B. S., Araújo L.M.M, Lins I.D., Moura M.D.C. & Droguett E.L. (2023), Prognostics and Health Management of Rotating Machinery via Quantum Machine Learning. *IEEE Access*, Vol. 11, pp. 25132-25151, doi: 10.1109/ACCESS.2023.3255417.

Nassif A.B., Talib M.A, Nasir Q. & Dakalbab F.M. (2021), *Machine Learning for Anomaly Detection: A Systematic Review*. IEEE Access, vol. 9, pp. 78658-78700 doi: 10.1109/ACCESS.2021.3083060.

Ochella S., Shafiee M. & Dinmohammadi F. Artificial intelligence in prognostics and health management of engineering systems (2022), *Engineering Applications of Artificial Intelligence*, Vol. 108, 104552, doi: 10.1016/j.engappai.2021.104552

Pang G., Shen C, Cao L.& Van Den Henge, A (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*. Vol. 54. Issue 2-38 pp 1-38 doi:10.1145/3439950

Sutton R.S & Barto A.G. (2018). *Reinforcement Learning. An Introduction*. The MIT Press.

BIOGRAPHIES

Miguel A. Sanz-Bobi is currently a Professor with the Computer Science Department, and also a Researcher with the Institute for Research and Technology (IIT), both within the Engineering School, Comillas Pontifical University, Madrid, Spain. He shares his time between teaching and research in topics related to the artificial intelligence field applied to diagnosis and maintenance of industrial processes. He has been the main researcher in an important number of industrial projects related to the diagnosis of industrial processes, incipient detection of anomalies based on models, knowledge acquisition and representation, and reliability and predictive

maintenance. All these projects have been based on a combination of artificial intelligence, new information technologies, and machine learning techniques.

Sarah Orbach is currently a student in the Engineering School CentraleSupelec. She mainly studied physics and computer science. During a gap year, she did two internships. The first one in collaboration to an Open Source web development project and the second in artificial intelligence with the Institute for Research and Technology (IIT) within the Engineering School, Comillas Pontifical University, Madrid, Spain. She is now specializing in bioengineering.

F. Javier Bellido-Lopez is an Electrical and Automatic-Electronic engineer from the Polytechnic University of Madrid (UPM). He is currently studying Physics and is a Researcher with the Institute for Research and Technology (IIT) of the ICAI Engineering School, Comillas Pontifical University, Madrid, Spain. His areas of interest include the application of Artificial Intelligence techniques to the monitoring and diagnosis of industrial processes, Data Analysis, Machine Learning.

Antonio Muñoz San Roque is currently a Professor with the Electronics and Communications Department, and also a Researcher with the Institute for Research and Technology (IIT), both within the ICAI Engineering School, Comillas Pontifical University, Madrid, Spain. His areas of interest include the application of Artificial Intelligence techniques to the monitoring and diagnosis of industrial processes, Time series forecasting, Machine Learning, and Electricity markets analysis.

Daniel González-Calvo is currently responsible for data-driven maintenance in the centralised maintenance unit at Iberia at ENEL/ENDESA. He obtained his master's degree in industrial engineering from the University of La Laguna and his PhD in industrial engineering (industrial doctorate) from the same university. He has worked on data projects for insular power generation systems and research on related data analysis techniques. His scientific and technical work has resulted in several publications and conferences. His areas of interest include predictive maintenance, industrial process optimisation and artificial intelligence applied to the energy sector.

Tomás Alvarez-Tejedor is currently Head of Thermal Maintenance Iberia at ENEL/ENDESA. He obtained his BSc, PhD degree in Engineering and MBA - Master in Business Administration at the Polytechnic University of Madrid (Spain) and his MSc - Master Science in The Gas Turbine Engineering Group at Cranfield University (UK). He has been working in the Spanish Electricity Market for more than thirty years and his background covers R&D projects on Advanced Power Generation Systems, Power Generation Asset Management and Combined Cycle and Gas Turbine Technology. His scientific and technical work are summarized in more than one hundred technical publications and conferences (EPRI, ASME, ETN, PowerGen,..etc). His areas of interest include the application of Artificial Intelligence techniques to power generation asset management.

Contrastive Metric Learning Loss-Enhanced Multi-Layer Perceptron for Sequentially Appearing Clusters in Acoustic Emission Data Streams

Oualid Laiadi[1], Ikram Remadna[2,3], El yamine Dris[1], Redouane Drai[1], Sadek Labib Terrissa[2], and Noureddine Zerhouni[4]

[1] *Research Center in Industrial Technologies (CRTI), Cheraga, P.O. Box 64, Algiers 16014, Algeria*
oualid.laiadi@gmail.com

[2] *LINFI Laboratory, University of Biskra*

[3] *National School of Artificial Intelligence (ENSIA) Algiers, Algeria*

[4] *FEMTO-ST Institute, Université Bourgogne Franche-Comté, CNRS, ENSMM*

ABSTRACT

Conventional structural health monitoring methods for interpreting unlabeled acoustic emission (AE) data typically rely on generic clustering approaches. This work introduces a novel approach for analyzing sequential and temporal acoustic emission (AE) data streams by enhancing a Multi-Layer Perceptron (MLP) with a contrastive metric learning loss function (MLP-CMLL)and Time Series K-means (TSKmeans) clustering. This dual approach, MLP-CMLL with TSKmeans, is crafted to refine cluster differentiation significantly. This method is designed to optimize cluster differentiation, bringing similar acoustic patterns closer and distancing divergent ones, thereby improving the MLP's ability to classify acoustic events over time. Addressing the limitations of traditional clustering algorithms in handling the temporal dynamics of AE data, MLP-CMLL with TSKmeans approach provides deeper insights into cluster formation and evolution. It promises enhanced monitoring and predictive maintenance capabilities in engineering applications by capturing the complex dynamics of AE data more effectively, offering a significant advancement in the field of structural health monitoring. Through experimental validation, we apply this method to characterize the loosening phenomenon in bolted structures under vibrations. Comparative analysis with two standard clustering methods using raw streaming data from three experimental campaigns demonstrates that our proposed method not only delivers valuable qualitative information concerning the timeline of clusters but also showcases superior performance in terms of cluster characterization.

Keywords: acoustic emission (AE), sequentially appearing clusters, data streams, structural health monitoring, contrastive metric learning, multi-layer perceptron (MLP)

1. INTRODUCTION

Structural Health Monitoring (SHM) is essential to ensuring the safety, longevity, and efficient maintenance of engineering structures across civil, mechanical, and aerospace fields. This discipline employs advanced technologies to proactively detect and address damages, aiming to avert catastrophic failures and optimize maintenance efforts. Among various SHM applications, the precision monitoring of bolted connections is particularly critical, given its profound impact on the structural integrity and stability of significant constructions such as bridges, aerospace structures, and wind turbines (Bolognani et al., 2018).

The vulnerability of bolted connections to loosening—and the profound implications of such—was dramatically underscored by the 2015 collapse of a 129-meter wind turbine in Sweden (Swedish Accident Investigation Authority, 2017). This incident, attributed to bolt looseness, resulted not only in significant financial loss but also highlighted the urgent need for early detection systems to prevent such disasters. While traditional bolt inspection techniques are effective, they are notably labor-intensive and can significantly interrupt operational workflows. This has led to a shift toward non-destructive testing (NDT) methods (Hoła & Sadowski, 2022), particularly the use of acoustic emission (AE) sensors (Sun, Yang, Li, & Xu, 2023; P. Xu, Zhou, Liu, & Mal, 2021; D. Xu, Liu, Li, & Chen, 2019), as more efficient alternatives.

AE sensors are distinguished by their ability to detect stress-induced changes within materials, offering a sophisticated means of identifying potential damages or loosening. Research, such as that conducted by Wang et al. (Wang, Song, Wang, & Li, 2013), demonstrates a correlation between AE signal energy and the axial load of bolts, enabling precise detection of bolt looseness through analysis of energy dissipation and signal amplitude. However, AE signals' complexity, marked by significant variations in amplitude and energy, coupled with susceptibility to environmental noise and interference, poses a significant challenge (Fu, Zhou, & Guo, 2023). Relying solely on a single AE characteristic often falls short in accurately reflecting bolt tightness. Therefore, there's a pressing need to develop innovative methods capable of quantifying AE signals' nonlinear characteristics and accurately interpreting bolt looseness, underscoring the demand for advanced analytical techniques.

The vast quantities of AE signals within data streams present a significant challenge in identifying ground truth, rendering supervised learning methods impractical for AE data interpretation or anomaly detection (Ramasso, Denoeux, & Chevallier, 2022; Ramasso, Placet, & Boubakar, 2015). This necessitates a pivot towards unsupervised learning techniques, such as K-means, fuzzy C-means (FCM), and Gaussian Mixture Models (GMM), to extract actionable insights from AE data. Among these approaches, Gaussian Mixture Models sequentially (GMMSEQ), introduced by Emmanuel Ramasso et al. (Ramasso, Denoeux, & Chevallier, 2022), stands out by incorporating temporal dynamics into the clustering of unlabeled AE data, thereby significantly enhancing parameter estimation related to damage progression.

Recent advancements highlight the growing significance of unsupervised and self-supervised learning methods, with a notable focus on contrastive metric learning. This approach harnesses the inherent similarities and contrasts within data to facilitate learning without the necessity for explicit labels, marking a pivotal shift toward more efficient representation learning (Saunshi, Plevrakis, Arora, Khodak, & Khandeparkar, 2019). By comparing input samples and manipulating their representations within the embedding space—drawing similar samples closer and distancing dissimilar ones—contrastive representation learning streamlines the learning process. It sidesteps the conventional need for labeling each sample, instead utilizing a pre-established similarity distribution to classify inputs into positive or negative pairs (Hassani & Khasahmadi, 2020).

Building on these insights, we propose a novel method that leverages the power of a Multi-Layer Perceptron (MLP) enhanced with a contrastive metric learning loss (MLP-CMLL), to adeptly handle AE data streams, particularly those exhibiting sequentially appearing clusters. The proposed MLP-CMLL approach, rooted in the principles of contrastive metric learning, aims to differentiate between similar and dissimilar features within the AE data, generating robust feature embeddings without the need for explicit labels. These embeddings serve as a powerful foundation for clustering, enabling our system to dynamically identify and group sequentially appearing clusters of AE data. By applying time series k-means clustering algorithm (TSKMean) (Huang et al., 2016), we can effectively cluster AE events based on both their feature similarities and their temporal characteristics. This integration enables the detection of sequentially appearing clusters, a common occurrence in AE data streams, thereby providing deeper insights into the material's behavior and the efficacy of the monitoring system.

The remainder of this paper is organized as follows: Section 2 introduces the proposed MLP-CMLL method, along with their respective data preprocessing methods. Section 3 describes the dataset and provides an analysis of experimental results. Finally, the main findings of this study are summarized in Section 4 along with a description of future work perspectives.

2. Proposed method

In this section, we delineate the architecture of the proposed framework, which aims to classify bolt tightening levels through the analysis of acoustic emission data streams. Figure 1 illustrates the overarching architecture of our proposed approach, specifically designed for clustering bolt tightening levels based on acoustic emission data streams. Subsequently, we will elaborate on the intricacies and functional components of the proposed framework, detailing each block's contribution to the overall system.

2.1. AE signal Preprocessing & Feature Extraction

The preprocessing and feature extraction of AE signals is a critical step in analyzing the raw data stream, performed through a three steps, initially outlined in (Kharrat, Ramasso, Placet, & Boubakar, 2016). The process begins with the data stream undergoing an initial filtration stage, employing a fifth-order high-pass filter with a cutoff frequency of 10 kHz and a passband ripple of 0.2 dB, effectively eliminating the DC component from the data.

- **Step 1**: **Wavelet filtering** Utilizing wavelet denoising on 250,000 sample frames achieves an optimal balance between computational efficiency and denoising quality. The chosen Daubechies "dB45" wavelet, featuring 90 coefficients and 14 decomposition levels, effectively identifies AE signal onsets (Kharrat et al., 2016). This step includes applying the soft Donoho-Johnstone universal threshold to the wavelet coefficients and adjusting for level-dependent noise, alongside correcting for any group delay introduced by the filtering process. Figure 2 displays the raw signal and denoised signal.

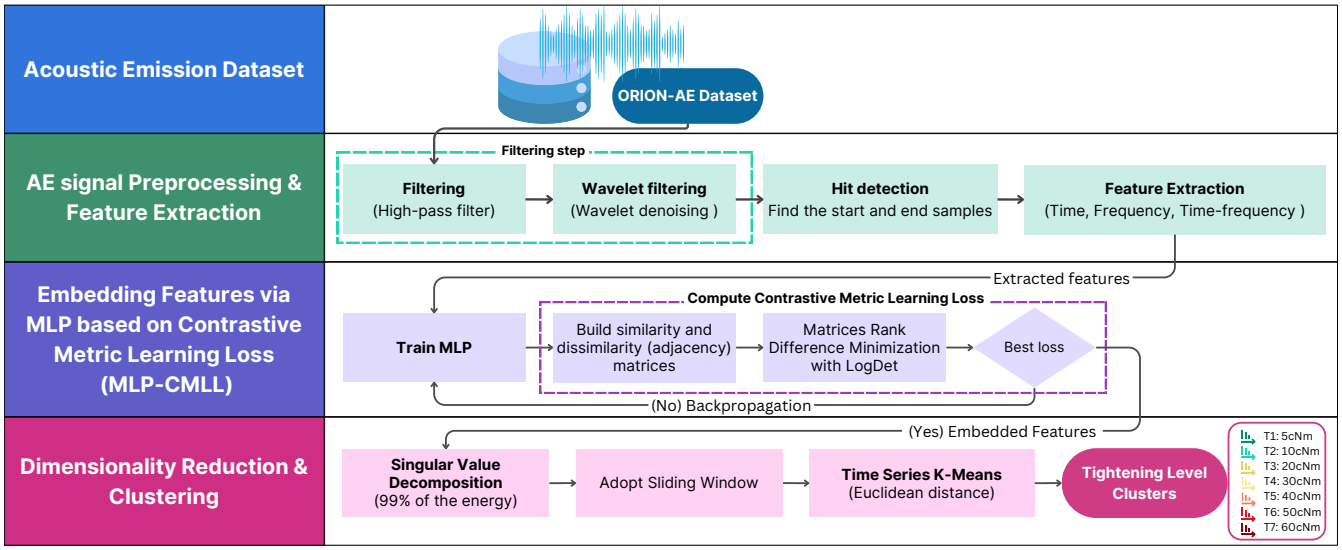


Figure 1. The general architecture of the proposed approach for bolt tightening level clustering.

- **Step 2**: **Hit Detection Procedure** aims to identify the start and end of each AE signal post-filtering based on amplitude thresholds (1.2 mV in this case). This step ensures that only relevant AE events are selected for analysis, utilizing specific counters ("HDT" 1100 μs and "HLT" 80 μs) to accurately demarcate signal boundaries (Kharrat et al., 2016).

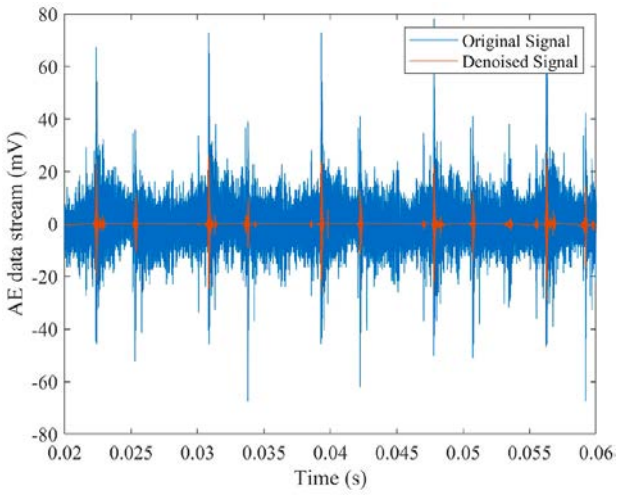


Figure 2. Raw signal and denoised signal.

- **Step 3 Feature Extraction**: Each detected AE signal is then analyzed to extract an extensive set of features, encompassing time-based and frequency-based characteristics (Kharrat et al., 2016; Sause, Gribov, Unwin, & Horn, 2012; Gonzalez Andino et al., 2000) such as rise time, counts, PAC-energy, amplitude, frequency metrics, signal strength, and energy distributions across specified frequency intervals. Figure 3 shows an AE signal and some typical features. Additional features include the Renyi number from the scalogram analysis using a Morlet wavelet and the frequency at maximum energy, providing a detailed signal characterization suitable for further analysis.

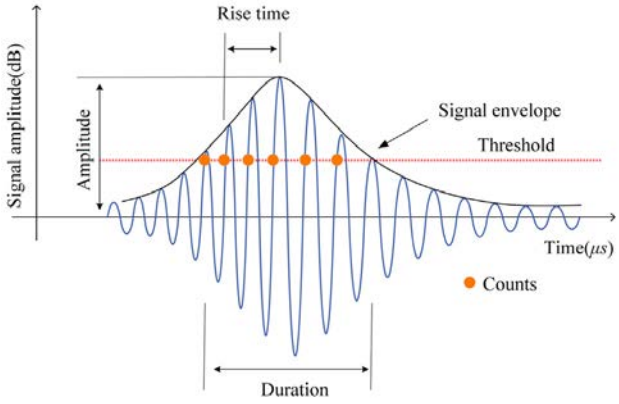


Figure 3. AE signal and some typical characteristics.

2.2. Embedding Features via MLP based on Contrastive Metric Learning Loss (MLP-CMLL)

This subsection will describe the proposed MLP based on the Contrastive Metric Learning Loss (MLP-CMLL) method. In the following, we will mention the details of our algorithm for learning a best loss metric based on an unsupervised metric learning with unlabeled data. The proposed contrastive metric learning framework is based on the combination of two methods, unsupervised EASE metric learning (Zhu & Koniusz, 2022a) and Generalized Laplacian Eigenmaps (Zhu & Koniusz, 2022b).

Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be an unlabeled AE data of n samples and m features. We propose a new MLP based on Contrastive Metric Learning Loss (MLP-CMLL) framework for unsupervised network embedding. To compute the loss function of the MLP framework, we calculate the logdet of scatter matrices based on the similarity and the dissimilarity (adjacency):

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \operatorname{Rank}\left(\mathbf{S}_{sim}\left(\mathbf{X}\right)\right) - \operatorname{Rank}\left(\mathbf{S}_{dis}\left(\mathbf{X}\right)\right) \quad (1)$$

Eq. (1) aims to compute a metric loss $\mathbf{\Theta}$ for each epoch that maximizes the similarity between similar features and minimizes the dissimilarity between dissimilar features.

Let:
$\mathbf{S}_{dis} = f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L}_{dis} f_{\mathbf{\Theta}}(\mathbf{X})$ and $\mathbf{S}_{sim} = f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L}_{sim} f_{\mathbf{\Theta}}(\mathbf{X})$
Then the LogDet relaxation becomes:

$$\begin{aligned}\mathbf{\Theta}^* &= \arg\min_{\mathbf{\Theta}} \log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L}_{sim} f_{\mathbf{\Theta}}(\mathbf{X})\right) \\ &\quad - \log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L}_{dis} f_{\mathbf{\Theta}}(\mathbf{X})\right) \\ &= \arg\min_{\mathbf{\Theta}} \log\det(\mathbf{I} + \alpha \mathbf{S}_{\text{sim}}) - \log\det(\mathbf{I} + \alpha \mathbf{S}_{\text{dis}})\end{aligned} \quad (2)$$

where $\mathbf{I}$ ensures $\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L} f_{\mathbf{\Theta}}(\mathbf{X}) > 0$ as $f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L} f_{\mathbf{\Theta}}(\mathbf{X})$ may be $\mathbb{S}_+^m$ leading to $\det\left(f_{\mathbf{\Theta}}(\mathbf{X})^\top \mathbf{L} f_{\mathbf{\Theta}}(\mathbf{X})\right) = 0$. Thus, we use $\log\det(\mathbf{I} + \alpha\mathbf{S})$ as a smooth surrogate for $\operatorname{Rank}(\mathbf{S})$.

$$\begin{aligned}\mathbf{L}_{\text{sim}} &= \mathbf{I} - \tilde{\mathbf{A}}_{\text{sim}} \in \mathbb{S}_+^n, \\ \mathbf{L}_{\text{dis}} &= \mathbf{I} - \tilde{\mathbf{A}}_{\text{dis}} \in \mathbb{S}_+^n.\end{aligned} \quad (3)$$

Let us also define normalized graph Laplacian matrices in Eq. (2) as in Eq. (3). Let $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \tilde{\mathbf{A}}$ and $\mathbf{D} = \operatorname{diag}\left(d_1, \cdots, d_n\right)$, where $d_i = \sum_j \mathbf{A}_{ij}$. We explain how we obtain $\mathbf{A}_{sim}$ and $\mathbf{A}_{dis}$ later in the text. From equations Eq. (3) and Eq. (2) we have:

$$\begin{aligned}\mathbf{L}_{\text{sim}} - \mathbf{L}_{\text{dis}} &= (\mathbf{I} - \tilde{\mathbf{A}}_{\text{sim}}) - (\mathbf{I} - \tilde{\mathbf{A}}_{\text{dis}}) \\ &= \tilde{\mathbf{A}}_{\text{dis}} - \tilde{\mathbf{A}}_{\text{sim}},\end{aligned} \quad (4)$$

As $\mathbf{L}_{\text{sim}} - \mathbf{L}_{\text{dis}} = \tilde{\mathbf{A}}_{\text{dis}} - \tilde{\mathbf{A}}_{\text{sim}}$, we obtain:

$$\begin{aligned}\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} &\log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{dis}} f_{\mathbf{\Theta}}(\mathbf{X})\right) \\ &- \log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{sim}} f_{\mathbf{\Theta}}(\mathbf{X})\right),\end{aligned} \quad (5)$$

where $\mathbf{A}_{\text{sim}}$ and $\mathbf{A}_{\text{dis}}$ are two different measurements with the opposite effect. Thus, we introduce parameter $\alpha > 0$ to balance the impact of these both terms.

Dissimilarity Matrix. Although one might design a linear projection based on the similarity relationship alone, we use both the dissimilarity information and the similarity matrix for learning a metric loss. Intuitively, in the context of a K-clustering task with n unlabeled samples and M_i queries for each cluster, we are addressing a problem where ($\mathrm{n} = \mathrm{K} \times \mathrm{M_i}$) samples are to be clustered into K groups. Here, off-diagonal entries are understood to signify distinct entities, whereas on-diagonal entries indicate identical entities. Thus, we form a dissimilarity matrix as the adjacency matrix of a densely connected graph:

$$\mathbf{A}_{dis} = \frac{1}{n}\mathbf{e}\mathbf{e}^\top - \mathbf{I}, \quad (6)$$

where $\mathbf{e}$ is an (n)-dimensional all-ones vector and $\mathbf{I}$ is the identity matrix.

Similarity Matrix. To measure the similarity between the pairs of samples, one has to choose a distance (or similarity measure) that will perform well in the clustering setting.

The typical choice for the measure of similarity is the RBF function $Z_{ij} = \exp\left(-\left\|f_\theta\left(\mathbf{x}_i\right) - f_\theta\left(\mathbf{x}_j\right)\right\|_2^2 / \sigma\right), \sigma > 0$ but the RBF function alone does not capture the structure of data. In this work, we claim that for the K-cluster learning task, the expected similarity matrix should be a K-block diagonal matrix. However, the similarity matrix based on the RBF kernel has no blockdiagonal structure.

Low-Rank Representation (LRR) (Liu, Lin, & Yu, 2010) expresses each data point $\mathbf{x}_i$ as a linear combination of other points, $\mathbf{x}_i = \sum_{j \neq i} Z_{ij}\mathbf{x}_j$, and uses the representational coefficient $\left(|Z_{ij}| + |Z_{ji}|\right)/2$ to measure the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. LRR takes the correlation structure of data into account, and finds a low-rank representation instead of a sparse representation. In this work, the LRR is applied in the following rank minimization problem:

$$\arg\min_{\mathbf{Z}} \left\|f_\theta(\mathbf{X}) - f_\theta(\mathbf{X})\mathbf{Z}\right\|_F^2 \quad \text{s.t. } \operatorname{rank}(\mathbf{Z}) = K. \quad (7)$$

Eq. (7) is solved in two stages: 1) $\mathbf{Z} = \mathbf{V}^\top\mathbf{V}$, where $\mathbf{V}$ is obtained from the skinny SVD of $f_\theta(\mathbf{X}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, and 2) for each row of $\mathbf{V}$, one only keeps top-K absolute largest entries of $\mathbf{\Sigma}$. Given the feature matrix $f_\theta(\mathbf{X})$, we obtain the representation matrix $\mathbf{Z}$ by solving Eq. (7). The similarity matrix is defined as $\mathbf{W}_{\text{sim}} = |\mathbf{Z}| - \operatorname{diag}(|\mathbf{Z}|)$.

We provide our implementation in Alg. 1. The proposed algorithm targets unsupervised network embedding by employing contrastive metric learning loss to enhance similarity among similar features while reducing dissimilarity among different ones. Central to this framework are the LogDet relaxation and Low-Rank Representation (LRR), both aimed at achieving an embedding that accurately captures the inherent structure of unlabeled data. This structured approach outlines a comprehensive step-by-step methodology for implementing the MLP-CMLL method, specifically designed to optimize metric learning loss in scenarios involving unlabeled datasets.

2.3. Time Series K-Means for clustering

Following the generation of high-dimensional embedded features via MLP-CMLL, with each feature vector comprising

Algorithm 1 MLP based on Contrastive Metric Learning Loss (MLP-CMLL)

Input: $\mathbf{X} \in \mathbb{R}^{n \times m}$: Unlabeled AE data of n samples and m features.
Initialize: Predefined MLP architecture, $\alpha > 0$.

Compute Similarity and Dissimilarity Matrices based on Laplacian Matrices:
$\mathbf{S}_{\text{dis}} = f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{dis}} f_{\mathbf{\Theta}}(\mathbf{X})$
$\mathbf{S}_{\text{sim}} = f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{sim}} f_{\mathbf{\Theta}}(\mathbf{X})$

Optimization:
while not converged **do**
    Solve for $\mathbf{\Theta}^*$ minimizing:

$$\begin{aligned}\mathbf{\Theta}^* &= \arg\min_{\mathbf{\Theta}} \log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{dis}} f_{\mathbf{\Theta}}(\mathbf{X})\right) \\ &\quad - \log\det\left(\mathbf{I} + \alpha f_{\mathbf{\Theta}}(\mathbf{X})^\top \tilde{\mathbf{A}}_{\text{sim}} f_{\mathbf{\Theta}}(\mathbf{X})\right) \\ &= \arg\min_{\mathbf{\Theta}} \log\det(\mathbf{I} + \alpha \mathbf{S}_{\text{sim}}) - \log\det(\mathbf{I} + \alpha \mathbf{S}_{\text{dis}})\end{aligned}$$

    Update MLP parameters.

    Adjust Matrices ($\mathbf{S}_{\text{sim}}$ and $\mathbf{S}_{\text{dis}}$) Based on MLP-embedded features.
end while

Output: MLP-embedded features $\mathbf{F}$ (transform $\mathbf{X}$ into feature-embedded space using best $f_{\mathbf{\Theta}}(.)$).

1024 dimensions, the next crucial step involves dimensionality reduction and the application of time series k-means for effective clustering. Singular Value Decomposition (SVD) (Wall, Rechtsteiner, & Rocha, 2003; Furnas et al., 2017) is employed to reduce the dimensionality of these embeddings, enhancing computational efficiency and preserves the essential characteristics of the embedded features.

Upon completing the dimensionality reduction, we employ a sliding window technique to integrate the time series k-means algorithm, a pivotal step for clustering AE data streams that exhibit temporal dependencies. This method involves segmenting the reduced feature set into overlapping windows, allowing for the dynamic nature of AE data to be captured over time. The sliding window approach (SW) organizes the data into sequences of a specified window size. We empirically choose the SW size as 50, with a step size (1) dictating the overlap between consecutive windows. This structuring is essential for maintaining the temporal continuity of AE events, facilitating the identification of clusters that evolve over time.

By applying time series k-means (Huang et al., 2016) to these windowed sequences, we can effectively cluster AE events based on both their feature similarities and their temporal characteristics. This integration enables the detection of sequentially appearing clusters, a common occurrence in AE data streams, thereby providing deeper insights into the material's behavior and the efficacy of the monitoring system.

3. Experimentation and Results

3.1. Acoustic emission dataset Description

The ORION-AE dataset (Ramasso, Verdin, & Chevallier, 2022) was obtained through a test rig known as ORION. The ORION is specifically designed to mimic the loosening phenomena commonly observed in bolted joints of structures in various industries, including aeronautics, automotive, and civil engineering. It is composed of two metallic plates linked together by three M4 bolts (as shown in Figure 4, enabling the simulation of bolt loosening under vibrational stress.

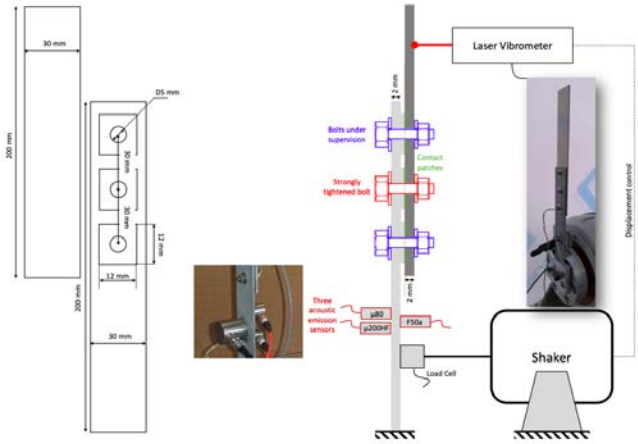


Figure 4. Setup description: part dimensions, sensors positions, bolts positions

The ORION-AE data are dynamically loaded with a vibration shaker and monitored with a laser vibrometer for velocity measurements and three AE sensors (micro80, F50A, micro200HF). The sensors sampled data at a rate of 5 MHz, producing datasets ranging from approximately 1.4 to 1.9 GB.

The ORION-AE dataset was generated by manually loosening a bolt on a test assembly and then subjecting it to 120 Hz harmonic vibrations, to simulate operational conditions. The experiment explored seven levels of bolt tightness (T1: 5cNm, T2: 10cNm, T3: 20cNm, T4: 30cNm, T5: 40cNm, T6: 50cNm, T7: 60cNm), with AE transients recorded for 10 seconds at each level. This procedure was repeated five times, resulting in five campaigns/datasets (B, C, D, E, and F), each with seven classes with 70 s of data for different sensors. Each campaign recorded varying numbers of signals, totaling 10,866; 9,461; 9,285; 15,628; and 17,810 signals, respectively. Note that, for campaign C, the level of bolt tightness 20 cNm is missing.

The seven tightening levels can be used as a ground truth when designing learning methods. This makes this dataset useful for developing and testing clustering and classification

methods for interpreting acoustic emission data.

For the purposes of this paper, analysis was focused exclusively on the micro-200-HF sensor, and only campaigns B, C and E were utilized to measure the performance of the clustering method. Figure 5 displays the tightening levels, acoustic emission and laser vibrometer data superimposed for measurements ”B” and sensor micro-200-HF (variable C).

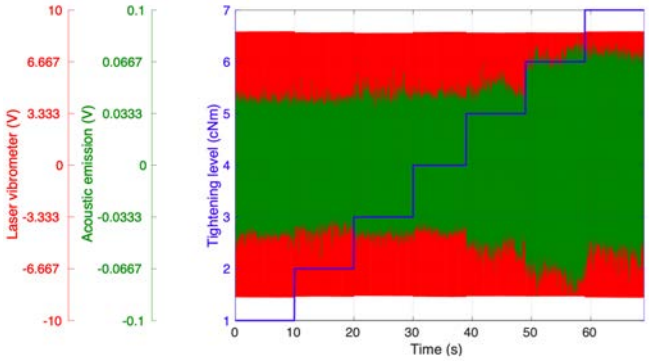


Figure 5. Tightening levels, acoustic emission and laser vibrometer data superimposed for measurements ”B” and sensor micro-200-HF (variable C).

3.2. Evaluation metrics

To properly evaluate the performance of clustering algorithms, such as TimeSeriesKMeans, on our test dataset, we use a variety of metrics. These metrics, as suggested by literature (Maulik & Bandyopadhyay, 2002), include:

- **Silhouette Score** evaluates cohesion within clusters and separation between them.
- **Davies-Bouldin Index** measures the average similarity between each cluster and its most similar cluster.
- **Adjusted Rand Index**, **Normalized Mutual Information (NMI)**, **Homogeneity**, **Completeness**, and **V-Measure** compare the clustering results to a ground truth, providing a measure of how well the clustering matches actual categories in the data.

3.3. Performance analysis

To demonstrate the effectiveness of the proposed unsupervised MLP-CMLL, we conduct numerous experiments to show the effectiveness of our embedded features compared to three different features, including raw data (AE signal Preprocessing & Feature Extraction), PCA (Kherif & Latypova, 2020) and SVD (Wall et al., 2003; Furnas et al., 2017). Tables 1, 2, and 3 show performance metrics for Campaigns B, C and E using different features over the TSKMeans cluster with sliding window.

Table 1. Performance metrics for Campaign B using different features over TSKMeans cluster with sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
RAW	0.818	0.296	**1.394**	0.842	0.844
SVD	0.818	0.296	**1.394**	0.842	0.844
PCA	0.818	0.296	**1.394**	0.842	0.844
MLP-CMLL	**0.875**	**0.335**	1.278	**0.884**	**0.884**

Table 2. Performance metrics for Campaign C using different features over TSKMeans cluster with sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
RAW	0.700	0.317	1.244	0.784	0.786
SVD	0.700	0.317	1.244	0.784	0.786
PCA	0.700	0.320	1.230	0.784	0.786
MLP-CMLL	**0.949**	**0.418**	**1.079**	**0.929**	**0.929**

Table 3. Performance metrics for Campaign E using different features over TSKMeans cluster with sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
RAW	0.738	0.228	**1.743**	0.800	0.805
SVD	0.738	0.228	**1.743**	0.800	0.805
PCA	0.738	0.228	**1.743**	0.800	0.805
MLP-CMLL	**0.854**	**0.300**	1.700	**0.866**	**0.867**

In the three tables 1, 2, and 3, the consistent outperformance of MLP-CMLL across all campaigns underscores the potential of sophisticated neural network-based feature extraction methods in enhancing clustering performance. It suggests that MLP-CMLL can adaptively learn and highlight the most relevant features for clustering, outpacing traditional dimensionality reduction techniques in capturing the essential structures of various datasets. Furthermore, the relatively close performance of SVD, PCA, and RAW methods across the campaigns might reflect their limitations in dealing with complex data structures or their potential redundancy when the raw data is already amenable to effective clustering. Analyzing three campaigns using various feature extraction methods within a TSKMeans cluster with a sliding window approach reveals consistent trends across performance metrics. The MLP-CMLL method consistently outperforms the other methods (RAW, SVD, PCA) in all evaluated metrics — Adjusted Rand Index (ARI), Silhouette score, Davies-Bouldin Index (DBI), Normalized Mutual Information (NMI), and Completeness—indicating superior clustering effectiveness. The RAW, SVD, and PCA methods display nearly identical performance across most metrics and campaigns, suggesting similar capabilities in handling clustering tasks. MLP-CMLL's higher scores across all metrics highlight its ability to capture more complex patterns and nonlinearities that linear methods might miss, resulting in better-defined and more accurately clustered data groups. This underlines the importance of method selection in data clustering to achieve optimal re-

Table 4. Performance metrics of different clustering methods for Campaign B using MLP-CMLL embedded features with (w/) and without (w/o) sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
GMM (w/o)	0.676	0.406	1.007	0.810	0.809
Kmeans (w/o)	0.623	**0.436**	1.000	0.786	0.795
MLP-CMLL+TSKMeans (w/o)	0.657	0.410	0.997	0.796	0.795
MLP-CMLL+TSKMeans (w/)	**0.875**	0.335	**1.278**	**0.884**	**0.884**

Table 5. Performance metrics of different clustering methods for Campaign C using MLP-CMLL embedded features with (w/) and without (w/o) sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
GMM (w/o)	**0.975**	0.420	1.050	**0.962**	**0.962**
Kmeans (w/o)	0.919	0.423	1.036	0.904	0.904
MLP-CMLL+TSKMeans (w/o)	0.919	**0.424**	1.037	0.904	0.905
MLP-CMLL+TSKMeans (w/)	0.949	0.418	**1.079**	0.929	0.929

sults based on specific campaign characteristics and objectives. Therefore, these observations suggest that while traditional methods like SVD and PCA have their merits, especially in contexts where computational simplicity and interpretability are key, advanced neural network-based approaches like MLP-CMLL offer a promising avenue for tackling more complex clustering challenges. Future work could explore further optimizations of the MLP-CMLL architecture, comparisons with other advanced machine learning techniques, and applications to a broader range of data types and clustering scenarios.

For Campaign B, MLP-CMLL shows the best performance across almost all metrics, highlighting its ability to extract meaningful embedded features that contribute to effective clustering. This suggests that the MLP-CMLL approach, with its presumably more nuanced understanding of the data structure, is particularly well-suited for the types of datasets represented in Campaign B. RAW, SVD, and PCA show similar performance in terms of ARI, Silhouette score, and other metrics. This could indicate that for Campaign B's dataset, the simpler dimensionality reduction techniques (SVD and PCA) do not provide significant advantages over using RAW data. This might be due to the nature of the data where the intrinsic data structure is either too complex for simple linear transformations to capture or perhaps is already in a form where raw data clustering is relatively effective.

For Campaign C, MLP-CMLL again outperforms other methods significantly in ARI and Completeness, reinforcing the value of advanced feature extraction methods in improving clustering outcomes. The improvement in the Silhouette score and DBI suggests that MLP-CMLL leads to more distinct, well-separated clusters than other methods. The performance gap between MLP-CMLL and other methods (SVD, PCA, and RAW) is notable, especially in terms of ARI and Completeness. This could imply that the Campaign C dataset contains complex patterns or high-dimensional structures that are better captured by the MLP-CMLL's feature extraction capabilities.

For Campaign E, MLP-CMLL's superiority is evident but less pronounced compared to Campaign C. It still leads in Adjusted Rand Index and Completeness, indicating its consistent effectiveness across different datasets. The similarity in performance between SVD, PCA, and RAW methods suggests that for Campaign E's data, the simple dimensionality reduction does not significantly impact the clustering performance, similar to Campaign B. However, the overall lower scores compared to Campaign B could indicate that Campaign E's dataset is inherently more challenging to cluster effectively, possibly due to noise, less distinct groupings, or more complex data structures.

Tables 4, 5, and 6 show performance metrics using different clustering methods for Campaigns B, C, and E with (w/) and without (w/o) sliding window.

Table 4 shows the performance metrics for Campaign B. The performance metrics for Campaign B provide a nuanced view of algorithm effectiveness. The Gaussian Mixture Model (GMM) showcases strong performance with an ARI of 0.676, suggesting a high degree of accuracy in clustering with respect to the true classifications. This is supported by an NMI of 0.810 and a Completeness score of 0.809, indicating a robust alignment between cluster assignments and actual data labels. The introduction of a sliding window with Time Series K-Means enhances its performance significantly, as evidenced by a jump in ARI to 0.875 and NMI to 0.884, underscoring the method's ability to capture temporal dependencies within the data. The Silhouette Score and DBI provide additional insights; despite a lower Silhouette Score (0.335) with the sliding window, indicating less clear separation between clusters, the method's overall effectiveness is not notably diminished, suggesting that the sliding window compensates by capturing temporal patterns not evident in spatial metrics alone.

Table 5 shows the performance metrics for Campaign C. Campaign C's analysis reveals the exceptional capability of the

Table 6. Performance metrics of different clustering methods for Campaign E using MLP-CMLL embedded features with (w/) and without (w/o) sliding window.

Method	ARI	Silhouette	DBI	NMI	Completeness
GMM (w/o)	0.626	0.311	1.453	0.716	0.718
Kmeans (w/o)	0.546	0.336	1.284	0.642	0.644
MLP-CMLL+TSKMeans (w/o)	0.598	**0.341**	1.371	0.671	0.673
MLP-CMLL+TSKMeans (w/)	**0.854**	0.300	**1.700**	**0.866**	**0.867**

GMM algorithm, achieving near-perfect ARI (0.975) and NMI (0.962) scores, which imply an almost flawless clustering outcome compared to true labels. This campaign highlights the impact of using a sliding window with Time Series K-Means, which achieves an ARI of 0.949 and an NMI of 0.929. These results suggest that the temporal structure captured by the sliding window significantly enhances clustering fidelity. The Silhouette Score (0.418 with the sliding window) and DBI (1.079 with the sliding window) indicate a balance between cluster cohesion and separation, affirming the effectiveness of incorporating temporal context in clustering analysis.

Table 6 shows the performance metrics for Campaign E. In Campaign E, the stark contrast in performance metrics between methods with and without sliding windows becomes even more pronounced. The use of the sliding window with Time Series K-Means propels its ARI to 0.854 and NMI to 0.866, suggesting a high degree of clustering accuracy that leverages temporal information effectively. Despite a lower Silhouette Score (0.300) with the sliding window, indicating potential overlap among clusters, the high NMI and Completeness scores (0.866 and 0.867, respectively) with the sliding window imply a successful capture of the intrinsic data structure. This campaign showcases the critical role of temporal analysis in clustering, especially for data where temporal patterns significantly influence the underlying structure.

Across campaigns B, C, and E, the analysis underscores the nuanced performance of GMM and Time Series K-Means, particularly when enhanced with a sliding window technique, across various clustering quality metrics. While simpler algorithms like Kmeans show competitive performance in specific metrics such as the Silhouette Score, the added complexity and temporal awareness of the sliding window modification in Time Series K-Means generally translate into superior clustering outcomes, especially in terms of aligning with true cluster structures and maintaining class completeness.

Advantages of MLP-CMLL Time Series K-Means (MLP-CMLL with TSKMeans)

From all tables 4, 5 and 6, the introduction of Contrastive Metric Learning Loss-Enhanced Multi-Layer Perceptron with Time Series K-Means (MLP-CMLL with TSKMeans) marks a significant advancement in clustering complex time-series data. This novel approach leverages the strength of contrastive learning to fine-tune the feature representation, significantly enhancing the clustering capability of TSKMeans by ensuring that similar instances are brought closer while dissimilar ones are distanced in the feature space. Our results underscore the efficacy of this method, particularly in achieving superior clustering performance metrics across all campaigns when compared to traditional approaches. Notably, the MLP-CMLL with TSKMeans exhibits remarkable improvements in metrics such as ARI and NMI, indicating not only an enhanced alignment with the true cluster structures but also a comprehensive capture of the intrinsic data relationships. This methodological enhancement introduces a powerful tool for time-series analysis, offering robustness against the challenges posed by the dynamic nature of temporal data and paving the way for more accurate, interpretable clustering solutions.

Our MLP-CMLL with TSKMeans vs. GMMSEQ (Ramasso, Denoeux, & Chevallier, 2022). In a comparative analysis between the novel MLP-CMLL+TSKmeans method and the GMMSEQ (Ramasso, Denoeux, & Chevallier, 2022) method across three experimental campaigns labeled B, C, and E, the performance is quantitatively measured using the Adjusted Rand Index (ARI). The ARI scores indicate the similarity between the clustering results and the true classifications, with a range from -1 to 1, where 1 denotes perfect agreement. For Campaign B, the MLP-CMLL+TSKmeans method significantly outperforms GMMSEQ, achieving an ARI of 0.875 compared to GMMSEQ's 0.772. This suggests a superior ability of the MLP-CMLL+TSKmeans to accurately match the true cluster structures. In Campaign C, both methods exhibit exceptional performance with MLP-CMLL+TSKmeans slightly leading (0.949 vs. 0.947), indicating that both are very capable but MLP-CMLL+TSKmeans shows a slight edge in capturing the clustering structure accurately. Campaign E again sees MLP-CMLL+TSKmeans outperforming GMMSEQ (0.854 vs. 0.799), reinforcing the method's robustness and accuracy in analyzing the complex dynamics of acoustic emission data streams. Overall, MLP-CMLL+TSKmeans consistently surpasses GMMSEQ in clustering performance across all campaigns, evidencing its effectiveness and the significant benefits it offers for structural health monitoring applications through better differentiation and handling of temporal dynamics within AE data.

4. CONCLUSION

This work introduced a new method for the analysis of acoustic emission (AE) data streams, which are inherently sequen-

tial and temporal. The study proposes a unique approach by enhancing a Multi-Layer Perceptron (MLP) with a contrastive metric learning loss function (MLP-CMLL) and time series kmeans, aiming to efficiently identify and analyze sequentially appearing clusters within the data. This novel loss function is meticulously designed to optimize the MLP by improving the differentiation between distinct clusters. The approach primarily concentrates on embedding sequences in a manner that clusters with similar acoustic patterns are brought closer together, while those with divergent patterns are distanced, thereby augmenting the MLP's capability to recognize and classify acoustic events based on their emission signatures over time.

The importance of this work lies in its ability to address the challenges associated with the precise characterization of dynamically forming clusters within AE data streams. Traditional clustering algorithms often falter in handling the temporal dynamics of AE data, where the sequencing and timing of events are crucial for a comprehensive understanding of the phenomena being monitored. By integrating a contrastive metric learning loss with an MLP architecture tailored to the specifics of sequentially appearing clusters in AE data streams, our method aims to unveil deeper insights into the formation and evolution of clusters. This approach promises to enhance monitoring and predictive maintenance in engineering applications by capturing the complex dynamics of AE data more effectively.

Through extensive experimentation and comparative analysis against conventional techniques, we validate the superiority of our proposed method in discerning the intricate dynamics of AE data. This work presents a robust analytical tool for the investigation of sequential clusters and their implications in the domain of structural health monitoring, offering significant advancements over existing methods in terms of cluster detection, characterization, and temporal analysis.

REFERENCES

Bolognani, D., Verzobio, A., Tonelli, D., Cappello, C., Glisic, B., Zonta, D., & Quigley, J. (2018). Iwshm 2017: Quantifying the benefit of structural health monitoring: what if the manager is not the owner? *Structural Health Monitoring*, *17*(6), 1393–1409.

Fu, W., Zhou, R., & Guo, Z. (2023). Automatic bolt tightness detection using acoustic emission and deep learning. In *Structures* (Vol. 55, pp. 1774–1782).

Furnas, G. W., Deerwester, S., Durnais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (2017). Information retrieval using a singular value decomposition model of latent semantic structure. In *Acm sigir forum* (Vol. 51, pp. 90–105).

Gonzalez Andino, S., Grave de Peralta Menendez, R., Thut, G., Spinelli, L., Blanke, O., Michel, C., & Landis, T. (2000). Measuring the complexity of time series: an application to neurophysiological signals. *Human brain mapping*, *11*(1), 46–57.

Hassani, K., & Khasahmadi, A. H. (2020). Contrastive multi-view representation learning on graphs. In *International conference on machine learning* (pp. 4116–4126).

Hoła, J., & Sadowski, Ł. (2022). *Non-destructive testing in civil engineering* (Vol. 12) (No. 14). MDPI.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, *367-368*, 1-13. doi: https://doi.org/10.1016/j.ins.2016.05.040

Kharrat, M., Ramasso, E., Placet, V., & Boubakar, M. (2016). A signal processing approach for enhanced acoustic emission data analysis in high activity systems: Application to organic matrix composites. *Mechanical Systems and Signal Processing*, *70*, 1038–1055.

Kherif, F., & Latypova, A. (2020). Principal component analysis. In *Machine learning* (pp. 209–225). Elsevier.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 663–670).

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(12), 1650–1654.

Ramasso, E., Denoeux, T., & Chevallier, G. (2022). Clustering acoustic emission data streams with sequentially appearing clusters using mixture models. *Mechanical Systems and Signal Processing*, *181*, 109504.

Ramasso, E., Placet, V., & Boubakar, M. L. (2015). Unsupervised consensus clustering of acoustic emission time-series for robust damage sequence estimation in composites. *IEEE Transactions on Instrumentation and Measurement*, *64*(12), 3297–3307.

Ramasso, E., Verdin, B., & Chevallier, G. (2022). Monitoring a bolted vibrating structure using multiple acoustic emission sensors: A benchmark. *Data*, *7*(3), 31.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., & Khandeparkar, H. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning* (pp. 5628–5637).

Sause, M. G., Gribov, A., Unwin, A. R., & Horn, S. (2012). Pattern recognition approach to identify natural clusters of acoustic emission signals. *Pattern Recognition Letters*, *33*(1), 17–23.

Sun, J., Yang, H., Li, D., & Xu, C. (2023). Experimental investigation on acoustic emission in fretting friction and wear of bolted joints. *Journal of Sound and Vibration*, *558*, 117773. doi: https://doi.org/10.1016/j.jsv.2023.117773

Swedish Accident Investigation Authority. (2017). *Windactionjvestas wind turbine collapse in lemnhult.*

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis* (pp. 91–109). Springer.

Wang, T., Song, G., Wang, Z., & Li, Y. (2013). Proof-of-concept study of monitoring bolt connection status using a piezoelectric based active sensing method. *Smart Materials and Structures*, *22*(8), 087001.

Xu, D., Liu, P., Li, J., & Chen, Z. (2019). Damage mode identification of adhesive composite joints under hygrothermal environment using acoustic emission and machine learning. *Composite structures*, *211*, 351–363.

Xu, P., Zhou, Z., Liu, T., & Mal, A. (2021). Determination of geometric role and damage assessment in hybrid fiber metal laminate (fml) joints based on acoustic emission. *Composite Structures*, *270*, 114068. doi: https://doi.org/10.1016/j.compstruct.2021.114068

Zhu, H., & Koniusz, P. (2022a). Ease: Unsupervised discriminant subspace learning for transductive few-shot learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9078–9088).

Zhu, H., & Koniusz, P. (2022b). Generalized laplacian eigenmaps. *Advances in Neural Information Processing Systems*, *35*, 30783–30797.

Applying Prognostics and Health Management to Optimize Safety and Sustainability at the First Adaptive High-Rise Structure

Dshamil Efinger, Giuseppe Mannone and Martin Dazer

Institute of Machine Components, University of Stuttgart, 70569 Stuttgart, Germany
dshamil.efinger@ima.uni-stuttgart.de giuseppe.mannone@ima.uni-stuttgart.de martin.dazer@ima.uni-stuttgart.de

ABSTRACT

Prognostics and Health Management (PHM) offers the potential to increase the acceptance of adaptive structures and to operate them in an optimal way. With suitable design and proper operation, adaptive high-rise structures enable significant increases in sustainability and service life extensions compared to passive high-rise buildings. The control loop for PHM provides a systematic overview of the contents related to PHM and their sequence. However, a framework is required for application to a complex adaptive system. Such a framework is presented in this paper. The framework is divided into the areas of system analysis and modeling as well as the PHM solution. A systematic approach is used to analyze the system and create the basis for full integration of all functional domains. This is then used in modeling to develop an adapted model structure. Finally, the PHM solution looks at the details of the approaches for diagnosis, prognosis, and health management.

1. MOTIVATION

The construction industry consumes significant resources and is responsible for a considerable proportion of CO_2 emissions (Thibaut Abergel & Dulac, 2018; OECD, 2015). Traditional load-bearing structures are typically designed for infrequent critical load cases. Additionally, there are numerous uncertainties, which are compensated in the design by safety factors. As a result, the load-bearing structure is often significantly oversized for most of its service life, which leads to increased resource demands and CO_2 emissions (Efinger et al., 2022). Actuators can be used to induce displacements, homogenize stresses in structures, and actively dampen vibrations. This enables the structural mass of such adaptive load-bearing structures to be reduced compared to passive structures by targeted reduction of cross-sections. In addition, the actuators need to be integrated into the structure and complemented by associated systems – including measurement systems, control systems and energy supply. Adaptive load-bearing structures offer the potential to save structural mass and – through suitable operation – to be more sustainable than conventional passive load-bearing structures (Efinger et al., 2022). At the same time, ensuring sustainability should not come at the expense of safety or serviceability. On the other hand, unnecessarily frequent maintenance measures need to be avoided for both sustainability and economic reasons. However, in addition to the parameter space for maintenance, there is also the parameter space for adaptivity control. This includes determining when which actuators exert how much force on the system and in what combination. As a result, stiffness and damping are controlled locally at the individual points, but also globally in the load-bearing structure. Depending on this, static and dynamic effects develop under the respective load, and more or less damage occurs in the individual elements of the structure. For both operation and maintenance, short, medium, and long-term objectives must also be considered and balanced. This high-dimensional problem cannot be solved with conventional methods for operating or maintenance strategies.

Prognostics and Health Management (PHM) offers the potential to improve the reliability and availability of technically complex systems in line with requirements. A comprehensive understanding of the application of PHM is crucial, especially when it comes to developing customized PHM solutions for complex systems.

The challenge in developing a universally applicable PHM solution lies in the high complexity and variability of the systems. Henß (Henß, 2021) highlights this problem and proposes a PHM control loop that offers a general approach to implementing PHM. However, this still needs to be embedded in the system for the application itself. To that end, this paper provides a framework that enables the practical implementation of the PHM solution and provides a structured approach for applying PHM to the complex system of an adaptive high-rise building.

To do this, a framework is presented that is specifically designed to apply PHM to complex systems. This framework

serves as both a method and a framework to enable effective implementation of PHM strategies. It is divided into two main parts: *System analysis and modeling* as well as *PHM-solution*.

In order to achieve this objective, section 2 – **basics** – initially introduces adaptive systems, the high-rise structure that serves as an example, and the PHM control loop.

Section 3, **system analysis and modeling**, focuses on understanding and modeling the specific features and behavior of the system under consideration. This phase is fundamental for the development of accurate predictive models and for the identification of relevant system parameters that are important for condition assessment and prediction.

This is followed by the **PHM-solution** in section 4, which comprises the development and implementation of solutions based on the findings and models of the system analysis. Specific techniques and procedures are used to implement system prognostic and the optimization of the system based on this.

Finally, section 5 gives a **summary**.

2. BASICS

This section introduces relevant basics for the adaptive system and introduces the PHM control loop.

2.1. Adaptive Systems

Systems that actively change with the help of sensors and actuators are referred to as adaptive systems. The system interacts with the environment and manipulates loads, for example, in such a way that load peaks are avoided (Sobek, Haase, & Teuffel, 2000). Adaptivity is enabled by a control process, whereby the input signals are fed into the system through sensors.

2.2. The D1244

The world's first adaptive high-rise building is called D1244. D1244 is a multi-functional experimental platform and is located on the campus of the University of Stuttgart. The load-bearing structure is activated by hydraulic actuators. Hydraulic pressure accumulators close to the cylinders ensure homogenization of the system pressure and minimize the switching requirements for the hydraulic pump. A central control unit and several module control units are available to control the hydraulic actuators, which are actuated by electro-hydraulic valves. Various sensors provide the controls with information on the load-bearing structure, the actuator system, and the ambient conditions. There are strain gauges for strain measurement in a redundant arrangement on the pillars and diagonals. LEDs mounted on the outer shell in the transition between the modules of the load-bearing structure serve as measuring points for an optical measuring system that uses cameras on two sides to record relative displacements and deformations of the load-bearing structure. The change in displacement of the actuators is recorded using displacement measuring systems on the hydraulic cylinders. There is a weather station on the roof of the building that records wind speed and wind direction.

2.3. Control of the Adaptive Structure D1244

The control loop contains the physical system, from which relevant variables such as stresses, deflection and other measured variables are recorded, a Kalman filter (KF) for condition monitoring and a linear-quadratic regulator (LQR) for controlling the adaptive components. Using the KF as an observer and estimator, the current system state $\hat{x}$ is estimated and iteratively transferred to the controller by a feedback loop. This allows unknown system variables to be determined and the system to self-adapt (Ostertag, 2021). Finite element models are used to investigate the equation of motion of the mechanical structure. The finite element method is used for the dynamic analysis of structures and the equation of motion of the structure. The discretization of the FE model at the nodal points results in the equation of motion according to (Ostertag, 2021; Gienger, Schaut, Sawodny, & Tarin, 2020):

$$M\ddot{q} + D\dot{q} + Kq = F_u u_{act} + F_\nu(\nu) \tag{1}$$

with the following boundary conditions:

$$\dot{q}(0) = \dot{q}_0, \;\; q(0) = q_0. \tag{2}$$

2.4. PHM Control Loop

The PHM control loop, as introduced by Henß (Henß, 2021), represents a systematic approach to optimizing the operation and maintenance of technically complex systems. Central to the control loop is the continuous interaction between the system and its optimization based on data from measurements, diagnosis, and prognosis. The optimization is fed directly back into the system, creating a closed loop that leads to continuous improvement.

The control loop can be abstractly divided into three main areas:

1. **Condition assessment:** This includes collecting data and analyzing it to determine the current system status.
2. **Forecast approaches:** Based on the condition assessment, a prediction is made about the future condition of the system.
3. **Optimization approach:** The results of the condition assessment and forecast are used to define and implement measures to improve the system.

Figure 1, based on (Henß, 2021), illustrates the PHM control loop including the three main areas formed. This illus-

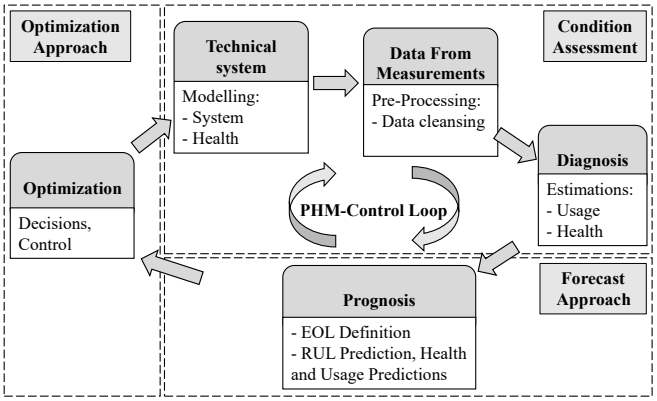


Figure 1. PHM control loop with its three main areas.

tration demonstrates the dynamic and cyclical nature of the PHM approach, which is based on continuous improvement and adaptation.

The condition assessment includes the evaluation of the system (system and health modeling), the evaluation of the data (data collection and cleansing) and the diagnosis (state estimation). This classification enables a precise assessment of the system and health statuses, based on which estimates can be made for the current state. Based on the estimates determined from the diagnosis, the predicted range within the condition forecast is determined. The downtime or, if an end-of-life definition has been selected, the remaining useful life (RUL) can be calculated. The basis for this is, among other things, the history to date. If information on future changes is available, these can also be considered. During optimization, the adjustable system settings are adjusted so that the objective function is as optimal as possible within the scope of the observation horizon. The information from the condition assessment and condition forecast is used for optimization. This holistic approach enables proactive maintenance and the optimization of operating processes, which leads to an extension of the service life and an increase in the efficiency of systems.

3. System Analysis and Modeling

This paper is built around the D1244 system, which offers ideal conditions for the application of PHM due to its complexity and adaptive capabilities. The special feature of the D1244 lies in the large number of influencing variables and the associated uncertainties. This makes the implementation of PHM a challenging task.

For a comprehensive integration of PHM into a complex system – such as an adaptive high-rise structure – a systematic approach is required. This is dealt with as part of the framework in this section. It is further subdivided into the steps of system analysis and modeling. During the system analysis in subsection 3.1, all aspects relevant for modeling the PHM application are determined. The models and their interrelationships are then built on this basis in subsection 3.2.

3.1. System Analysis

This subsection develops the content to create models for state estimation, prediction, and health management. A comprehensive system analysis is carried out for this purpose. This begins with the overall objective, which also provides target parameters, followed by the system description. In addition, the requirements and boundary conditions for operation and its optimization are extracted. From this, the dimensions of the system are derived and supplemented by further conditions. Influencing factors are then identified on this basis. Lastly, uncertainties for the target parameters and the boundary conditions can be derived from the target parameters and influencing factors.

Once these steps have been completed, all areas are available for developing models for the PHM application.

3.1.1. Objective

The objective of the PHM control loop according to Henß (Henß, 2021) is to operate and maintain the system in such a way that the target parameter is maximized, and the existing boundary conditions are met. A target parameter describes a measurable or countable variable from the objective. For the use case of the adaptive high-rise system this means:

Firstly, the environmental impact should be as low as possible, with the CO_2 equivalent being assessed. Secondly, increasing the service life of the system beyond a certain level may also be a further goal. Thirdly, this needs to be done in compliance with the requirements and boundary conditions.

There is now a conflict of interest between the goals of lowest possible CO_2 equivalence and increased service life. The lowest possible CO_2 equivalent would be achieved with the shortest possible service life and an increase in service life would presumably require additional CO_2 expenditure. For this reason, the CO_2 equivalent is used as a reference for its opposing benefit – i.e., the service life. This means that the CO_2 equivalent is normalized to the useful life. Nevertheless, it is possible that the target parameters compete. Therefore, an individual objective function needs to be formed for each use case from the existing target parameters and boundary conditions for multi-criteria optimization. This is implemented as part of the modeling in subsection 3.2.1.

3.1.2. System Description

Now that the objective is set, it is important to clearly define what it applies to. The system description serves to make the system and its components tangible and to separate the area under investigation from the environment through a defined system boundary. In addition to a physical and a signal-related system boundary, the system boundary also includes

a temporal dimension. For the example under investigation, this means:

The system of interest is the entire system of the adaptive high-rise structure, described in subsection 2.2. In addition to the load-bearing structure, this also include the functional domains of the actuators, various sensors, physical control, and transmission elements, the software control and the energy supply. Furthermore, maintenance with all its domains – including spare parts stocking, capacity planning, etc. – is also included in the system scope. Figure 2 shows the domains of the adaptive structure schematically and indicates their allocation. The temporal consideration starts from the beginning of the building usage until the end of the service life. The manner in which the adaptation function is carried out by the components of the adaptive system corresponds to the description in subsections 2.1 and 2.3. To do this, a functional adaptation function is required. If the adaptation function is to be performed and it is not functional, the adaptation function fails.

3.1.3. Requirements and Boundary Conditions

The first question that arises for a defined system or its optimization is that of the general requirements and boundary conditions. They are set externally or internally and need to be fulfilled. There are elementary requirements for a structure such as a high-rise building. These include the fact that they have to enable their intended use and that they have to have a certain geometrical shape. Standards such as the Eurocodes formulate further requirements. There are also project-specific, definable requirements that the building has to fulfill. The probability of partial or total failure of the structure or its function is of central importance. Suitable measures must be taken to ensure that the probability of occurrence is lower than the limit values defined as acceptable. Typical design measures for this are redundancies in the functional structure or the oversizing of relevant structural components. The limit values to be fulfilled are, on the one hand, limit values for the designed service life and, on the other hand, limit values for failure at any time. The Eurocodes typically work with two different safety levels for buildings. The stricter one deals with the risk of structural failure. Whereas the less stringent one concerns, for example, the occurrence of non-failure-critical cracks or the occurrence of building vibrations that could potentially be perceived as unpleasant by users. While the stiffness – and therefore the vibration behavior – of passive load-bearing structures cannot be actively changed, this is possible with adaptive structures. This means that compliance with building vibrations defined for user comfort could be linked to the presence of users in the building. In this way, the limit value would change over time – if there were people in the building, it would take effect, if there are no people in the building, it could be raised.

3.1.4. Dimensions

The existing dimensions can be developed based on the requirements and can be further enhanced based on the objective and the boundary conditions. To fulfill the requirements, the system has to perform functions, but these are not always requested in the same way.

For the D1244, this gives rise to the three existing dimensions – the function request dimension, the function availability dimension, and the temporal dimension. These are supplemented by additional criteria from the objective and the boundary conditions. From the objective, this are the target parameter of CO_2 equivalent and the extension of the service life. The boundary conditions provide the time-variable risk limit for the failure of the adaptation function.

The dimensions can be examined in more detail. The aspects to be found there can influence each other within the dimension as well as between the dimensions or are dependent on each other. For example, the temporal dimensions consist of endless gradations between the past, the present and the future. The dimension of function request is composed of the aspects of building usage, external influences such as wind and weather as well as the limits of safety levels.

However, the most comprehensive dimension is that of function availability. It can be divided into the physical system, non-adjustable and adjustable system properties, and system monitoring. The following lists show some of the contents.

Physical system:

- *Load-bearing structure:* Load capacity; Statics; Dynamics; Health
- *Actuators:* Dynamics; Fault; Failure; Health
- *Sensors:* Fault; Failure; Health
- *Physical control elements:* Dynamics; Fault; Failure; Health
- *Physical transmission elements:* Fault; Failure; Health
- *Energy supply:* Fault; Failure; Health

Non-adjustable system properties:

- *Function structure:* Redundancy; Compensation possibilities
- *Failure behavior individual elements:* Reliability; Health; Damage factor; Failure behavior

Adjustable system properties:

- *Control options:* Control limits; Control models; Actuators; Ph. control el.; Energy supply
- *Maintenance:* All physical domains; Spare parts; Tools; Personnel; Measures

System monitoring:

- *State detection:* Position; Force; Stress; Pressure; Motion; Acceleration; Control; Health; Maintenance; Usage; External influences; function request

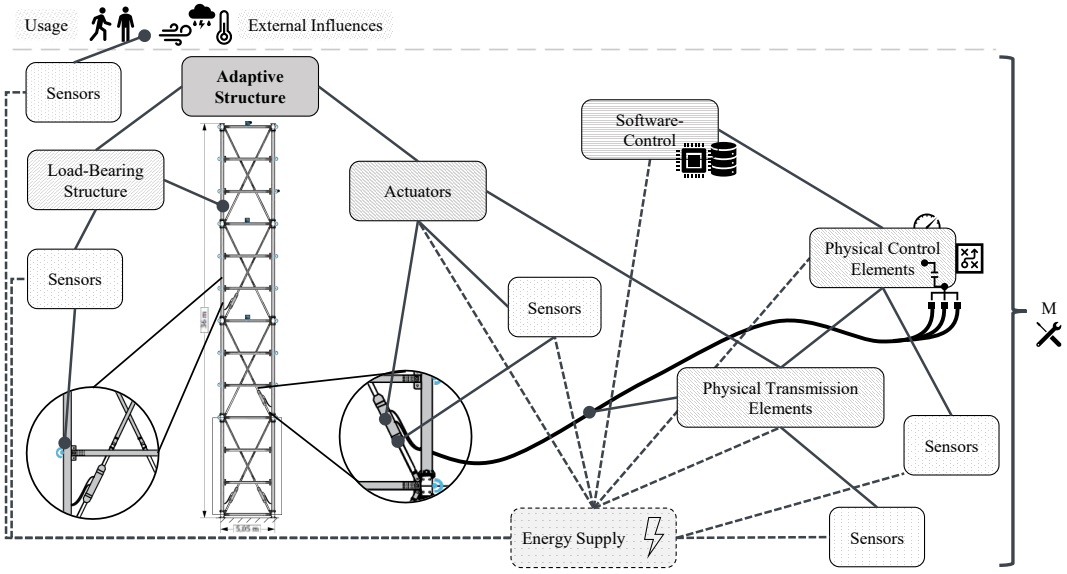


Figure 2. Schematic illustration of the system.

- *Fault detection:* All physical domains with exception of load-bearing structure; Control
- *Failure detection:* All physical domains; Control

Using the dimensions, the subsequent steps of the system analysis can be systematically examined for all existing aspects of the system. The dimensions can also be used to systematize the interconnection of aspects and dependencies in the functions for subsequent modeling. Aspects with identical themes can therefore be found in different dimensions.

3.1.5. Influencing Factors

This step collects all existing influencing factors. Influencing factor is anything that influences the fulfillment of the requirements, the boundary conditions, or the objective at any point in time. For the PHM application, the influencing factors needs to be divided into those that cannot be adjusted during operation, those that can only be adjusted indirectly and those that can be adjusted directly by PHM.

3.1.6. Uncertainties

Uncertainties describe a lack of or imprecise knowledge about something. To utilize the existing knowledge, it is necessary to quantify the uncertainty. Otherwise, unknown risks would be taken. These can be of an economic, social, or ethical nature. Uncertainties can be determined from the combination of target parameter and influencing factor as well as their relationship and classification in the existing dimensions.

Generalized uncertainties of the respective influencing factors from the following areas have to be taken into account. This applies to each measurement and model level and influences dependent model and system areas.

- Measurement uncertainties: Measurement errors, measurement data noise, etc.
- Model uncertainties: Model errors, model inaccuracies, etc.
- Stochastic effects
- Time

From the levels of the measured variables to the sub-models and the models, more influences of uncertainties are added. Some of these can be reduced using appropriate models to reduce uncertainties. An example of this is checking for measurement errors from sensor signals using a model that checks several sensor signals for plausibility and can perform error identification.

The temporal dimension plays a crucial role in uncertainties. In a present state, uncertainties from the areas of measurement uncertainty and model uncertainty can be present. If the corresponding data is available for the past, the situation is the same there. When considering future points in time, however, the uncertainty from the temporal development is always added. Figure 3 shows an uncertainty space that spans the areas of measurement and model uncertainties and the temporal dimension. There is no temporal uncertainty for the past and the present, provided that the data from the past is still fully available. For the future, however, the temporal uncertainty increases more with increasing temporal distance. The measurement uncertainty and the model uncertainty do not have a generalizable function from lower to greater uncertainty. They must be estimated individually for each case under consideration.

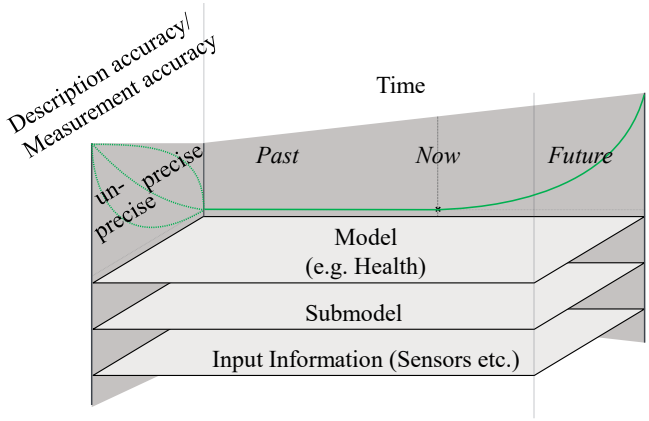


Figure 3. Illustration of a dimension model with dimensions of uncertainties.

3.2. Modeling

Subsection 3.2.1 establishes the target space and objective function based on results, requirements, and boundary conditions. The creation and interconnection of model structures across various levels are detailed in 3.2.2. System integration for the structural control of an adaptive high-rise structure is discussed in 3.2.3.

3.2.1. Target Space and the Objective Function

The target space represents the scope in which all specifications and requirements are met. It therefore describes permissible solutions. The objective function defines the weighting of the target parameters in relation to each other. Permissible solutions fulfill all requirements, the best possible solutions achieve the highest values for the objective function. In concrete terms, this means that the probabilities of partial or total failure of the structure or its function must be always kept below the limit values and for the entire planned service life of the structure. Some of the limit values can change over time, depending on factors such as building occupancy. At the same time, the overriding objective is to achieve the lowest possible CO_2 equivalent and possibly an increase in the service life of the building. The weighting of the two target parameters needs to be described by way of a relationship. Further boundary conditions and target parameters are possible and could be added in the same way. The optimization of such an objective function is not trivially solvable for such a complex system. Therefore, the use of fuzzy logic is proposed in section 4 in order to enable an assessment of all influencing factors based on the information about the system using the hybrid approach described there.

3.2.2. Model Structure

The development of the model structure relates to two designations of models and their structure. On the one hand, it is about the dimensional models and their interconnections. On the other hand, it is about functional models and their interaction.

Function Models The function models are divided into the three areas of condition assessment, forecast assessment and optimization approach introduced in subsection 2.4. Figure 4 also shows an overview of the structure of some function models for the PHM solution of the adaptive high-rise building. In accordance with the PHM control loop according to Henß and figure 1, the diagnosis in the condition assessment provides input for the prognosis in the forecast assessment and this for the optimization. Optimization continues until an accepted condition forecast is reached, for which the instruction to adjust the relevant influencing factors is given. In addition to these optimized adjustment instructions described above, there are also system variables that are also provided externally by the condition assessment from the PHM solution. The adjustment instructions contain parameters with different change behavior over time. For example, the instruction of a maintenance measure after several years of operation and with an execution time of several days or weeks. But also, the current state of stress on an element of the load-bearing structure or the pressure in one of the actuators.

- The **diagnostics** includes function models for the detection, determination or estimation of
  - Failures: for example of individual actuators, sensors, valves or load-bearing elements
  - Faults: for example in the measurements of sensors, the control behavior of valves or the control instructions of the control system
    For example, (Stiefelmaier, Böhm, Sawodny, & Tarín, 2023) provides a concept for detecting sensor or actuator faults that can be integrated.
  - System states: for example, for the individual parts of the load-bearing structure, actuators, sensors, feed and control systems, or maintenance processes in relation to their positions, mechanical or electrical stresses or hydraulic pressure, the dynamic state or the state of damage
- The **prognostics** contains function models for forecasting all states of the condition assessment. It is fed with information from the diagnostics. External predictions are also included. These include forecasts for wind and weather, personnel capacity, spare parts capacity and, if applicable, utilization and the planned remaining service life.
- **Optimization** takes place in Health Management. Function models are available to evaluate the optimized solution and to optimize the adjustable influencing factors.

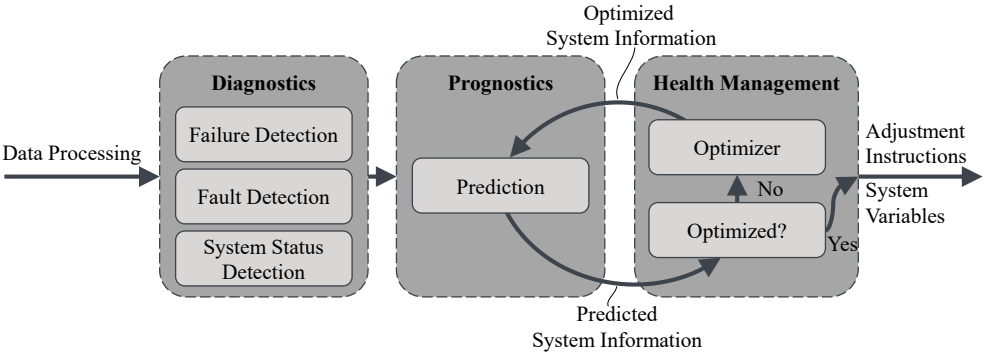


Figure 4. Some feature models of the PHM-solution.

Dimension Models The dimensional models or models of dimension aspects form different layers on top of the function models. They interact with each other and with the function models. Figure 3 shows such a dimension model using the example of health and taking the temporal dimension into account. The dimension model uses various input information, which is prepared and processed in sub-models before being merged in the dimension model.

The correlations can be illustrated using a simple example of a fault in the measured value of a sensor. In the case of fault detection in the measured value of a sensor, the function model for the fault detection is used to handle an influence from the dimension aspect of sensors. To assess whether or with which other sensor information an evaluation is possible, the model for the dimension aspect of function structure is used to work on redundancies or compensation options. These may be reduced in number by the results of the function model for failure detection. The temporal dimension may also be used to detect changes in the data history.

3.2.3. System Integration on the Example of Structural Control

This subsection uses the example of structure control to describe how system integration can be carried out with the PHM solution. To do this, we refer back to Figure 4 in the previous subsection. The PHM solution provides adjustment instructions. These must then be processed – for structural control in the control of the actuators. The system integration for structural control can be divided into two higher-level subareas if this is abstracted:

- Intelligent control model
- Intelligent monitoring and assessment model

The conventional approach from subsection 2.3 for controlling the high-rise demonstrator involves using a linear-quadratic controller in a closed control loop with feedback and a Kalman filter as an observer. This means that only a limited adaptation of the control behavior is possible.

In (Dakova, Heidingsfeld, Böhm, & Sawodny, 2022) the authors present an agent-based fatigue level controller through the use of a model predictive control (MPC) and the use of a cost function for the damage in the elements. This controller is adopted for structural control as a so-called score MPC. The PHM solution provides it with the control difference $e(t)$ and the optimized scores for each actuator as adjustment instructions. In this way, the PHM solution defines the intensity of the control and the distribution between the actuators in the system. The block diagram for the control loop is shown in Figure 5. The score MPC thus forms the intelligent control model, and the PHM solution forms the intelligent condition monitoring and evaluation model.

4. PHM-SOLUTION

This section covers the diagnostics and prognostics approach, detailing the use of a hybrid model and a health management agent that utilizes optimized system information for improved control behavior. It also introduces a data-driven strategy employing fuzzy neural networks (FNN) to manage the complexity of the D1244 system. This method facilitates the efficient analysis of extensive, multi-layered data, enhancing decision-making in maintenance and operations. Fuzzy neural networks enable the PHM solution to adapt and refine predictions, crucial for addressing challenges in complex systems like the D12444.

4.1. Diagnostics and Prognostics Approach

The effective assessment and prediction of the health of complex systems have requirements that cannot be met by conventional approaches. As described by Kim et al. (Kim, An, & Choi, 2017), Goebel et al. (Goebel et al., 2017), and Si et al. (Si, Zhang, & Hu, 2017), hybrid approaches offer a promising solution. These approaches combine data-driven algorithms with physical degradation models to improve the accuracy of condition assessment and prediction.

Such a hybrid approach is made up of various sub-models, each of which depicts individual dimensions of the system,

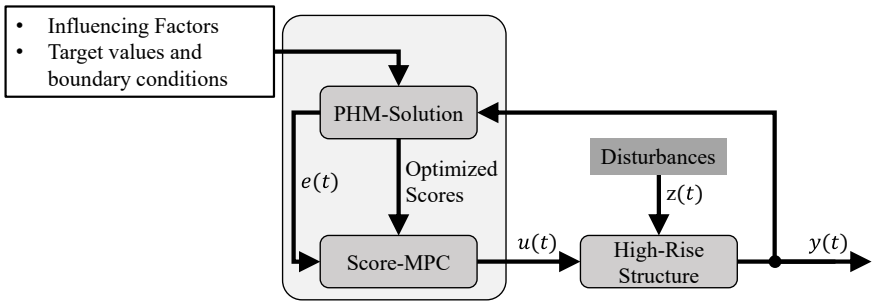


Figure 5. Control loop for the D1244 optimized for PHM application.

and which are linked together. Training and experience data are essential for the development and refinement of these approaches. This data includes not only the current measured values recorded by sensors, but also historical data that is used to calibrate and train the algorithms.

4.1.1. Design of the Hybrid Approach

The hybrid approach combines data-driven and physically based models to enable a comprehensive analysis of the failure behavior of system components. Pure data-driven models based on mathematical functions can only approximate reality without taking underlying physical models into account. The integration of a physical model, on the other hand, enables a deeper insight into the behavior of the system. This combination not only increases the accuracy of the predictions, but also expands the database for training the algorithms and promotes a holistic understanding of the overall system.

A schematic setup for the application of such an hybrid approach to a complex adaptive system, such as the D1244, is shown in Figure 6. Data- and physics-based methods are used to both estimate the current state and predict the future state.

A degradation model forms the basis for making valid statements about the system based on sensor data and minimizing inaccuracies. The combination of data-driven findings and physical models provides a robust basis for the reliable evaluation and prediction of system performance.

To optimize the informative value of this approach, continuously recorded training data is integrated into the prediction models over the system's service life. This enables a flexible and effective response to unexpected system changes.

This approach distinguishes between two sub-models, which are discussed in detail in the following sections. These submodels are the state estimation and the state prediction.

4.1.2. State Estimation

The state estimation serves as the basis and data foundation for the forecasting approaches. For this purpose, system and condition modeling is carried out and data from the technical system is recorded by sensors. The degradation model plays a key role here, as it provides a basis for the estimates. Degradation modeling is conducted using damage mechanisms on the components or estimates of the system modeling, for example using KF. The assessment of the condition takes place in the diagnosis: Through data, states that affect the use and health of a technical system are estimated in diagnostics based on collected information.

4.1.3. State Prediction

In addition to current sensor data, the hybrid approach can also incorporate training data from past measurements for the state prediction. This allows the forecast to be improved. This therefore represents a continuously learning process.

4.2. Health Management Agent

During optimization, the predicted system information is finally optimized iteratively, and a decision is made for the system. Optimization goals such as fault tolerance, sustainability and fault prevention are considered, for example to enable proactive and environmentally friendly maintenance planning. Based on the optimized system information, the control behavior of the system can be adapted and improved.

Calculating the score is therefore one of the most important tasks of optimization and can be achieved by using fuzzy methods in combination with NN. The use of FNN has the advantage that a more precise classification and assignment of the optimized system variables is possible and thus a more precise score calculation can be realized for the respective actuators.

4.2.1. Optimization of System Information

The optimization of the predicted system information as part of health management aims to improve the service life and added value over the entire life cycle by integrating optimization factors. This strategic adjustment affects both maintenance planning and the efficiency of use of the system components. Through targeted control, component functions can be adapted not only to extend the service life but also to

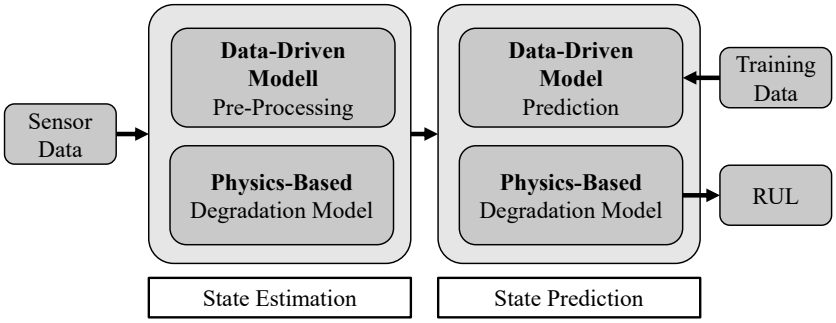


Figure 6. Schematic setup of the hybrid approach for a complex system.

prevent premature failures. The process is iterative and is based on the continuous analysis of the remaining useful life (RUL), which considers both risk factors and life cycle analysis (LCA) factors. This comprehensive approach enables dynamic adaptation of maintenance strategies and a flexible response to changing operating conditions or new findings about the condition of the components. By integrating both data-driven insights and physical model assumptions, a more precise forecast and thus more effective planning and control of maintenance measures is achieved.

The continuous improvement of the management system through this iterative approach not only supports an extended service life and increased reliability of the system components, but also helps to reduce maintenance costs and increase overall efficiency. This methodical approach is therefore an essential part of strategic maintenance and operational management, creating a balance between preventive maintenance and operational flexibility.

4.2.2. AI-Supported Optimization Using Fuzzy Neural Networks

At the core of the Prognostics and Health Management solution is the implementation of fuzzy neural networks, which use the hybrid approach to make adaptive decisions regarding the system components and adjust the control behavior accordingly. Embedding fuzzy logic (FL) in neural networks enhances their ability to precisely interpret and process complex system states and dynamics. FL mimics human thinking by making it possible to represent complex relationships in an understandable form. In addition, the use of FL enables a sophisticated analysis of uncertainties and ambiguities within the system data, allowing FNNs to make efficient decisions even in the presence of incomplete or fuzzy information. Fuzzy sets and rules are directly integrated into the structure of the neural networks, resulting in a synergetic architecture of fuzzy neural networks (de Campos Souza, 2020; Mishra, Sahoo, & Mishra, 2019).

The special feature of this approach lies in the improved interpretability of complex systems and the effective handling of data uncertainty. The use of FNNs enables a deep understanding and clear interpretability of system dynamics and state evaluations, which is crucial for the optimization of PHM solutions (de Campos Souza, Lughofer, & Guimaraes, 2021).

4.3. Fuzzy-Based Score Calculation

The interface between the PHM solution and the controller from 2.3 requires a control difference to be defined and transferred to the control system. For this purpose, a PHM solution is added to the control system, which calculates a score value for the individual actuators using extended optimization. The score calculation can be performed using various methods, including simple weighting or the application of complex logical operations using fuzzy rules in fuzzy neural networks (FNNs). The objective function from subsection 3.2.1 and the influencing factors from subsection 3.1.5 are included in the calculation of the score. For example, deferring maintenance to reduce CO_2 emissions can increase the risk of failure. All of these factors needs to be considered as part of an iterative optimization, as illustrated in Figure 4. The influencing factors can be integrated into a calculation function as weights. One implementation option is the formulation of fuzzy rules, whereby the weighting of the individual factors within the FNN is included in the evaluation. This allows complex logical relationships based on the data to be considered and the result to be interpreted as a score value.

The main objective of using an FNN is to make the score calculation for adaptive control efficient and effective. By integrating fuzzy rules, the various influencing factors are categorized and evaluated by fuzzification layers in the network. This allows not only a specific weighting of the factors, but also a flexible adjustment of these weightings within the layers. Training data from previous calculations or empirical values can also be used to train the network. A particular advantage of fuzzy networks is that they do not require an explicit model and that they can simplify the score calculation despite complex mathematical relationships through fuzzification.

4.3.1. Design of a Fuzzy Network

A fuzzy-based neural network consists of several core components. These are:

- **Fuzzy logic:** Enables the processing of uncertainties and ambiguities through the use of fuzzy sets and fuzzy operators.
- **Membership Functions:** Each fuzzy set is represented by a membership function, which specifies the degree to which an input value belongs to a set.
- **Fuzzy Rules:** Fuzzy rules, formulated as "IF-THEN" statements, define how input values should be processed based on their membership of different fuzzy sets.
- **General neural network structure:** Similar to traditional neural networks, including layers of neurons that are interconnected by weighted connections.

4.3.2. Definition of Fuzzy Rules

As the decisions of the adaptive system are based on the fuzzy rules, it is important to define the most precise boundaries possible when recording the rules. The rules should represent the characteristics of the training data sets as accurately as possible. To define precise fuzzy rules, approaches such as the improved Wand-Mendel method can be used. This tool makes it possible to determine fuzzy rules directly from the data sets. However, as this leads to further uncertainties, it is important to determine the fuzzy rules iteratively to keep the uncertainties as low as possible. And use methods such as the WM method as an additional option.

Since the determination of fuzzy rules is a continuous process, it is essential that the fuzzy rules within this framework can be adapted flexibly and adaptively to changes to be able to react effectively to modifications of the underlying rules.

4.3.3. Example Use Case of Fuzzy Neural Network

To demonstrate the practical application and effectiveness of FNNs across various adaptive control systems, we explore an example using the D1244. This use case illustrates how FNNs can be seamlessly integrated into complex environments to manage and improve system responses dynamically.

FNN Implementation: The FNN architecture in this scenario consists of several key layers, each tailored to handle the specifics of sensor data interpretation and decision-making in a fuzzy context:

- **Input Layer:** Directs raw sensor data into the network.
- **Fuzzification Layer:** Converts numeric sensor inputs into fuzzy values using membership functions. These functions define linguistic terms such as low, medium, high, which are easier to handle in rule-based logic processing.
- **Inference Layer:** Implements fuzzy logic rules that determine the control responses based on the fuzzy inputs. This layer combines the fuzzy terms using logical operators and forms the backbone of decision-making within the network.
- **Defuzzification Layer:** Converts the fuzzy conclusions back into precise control outputs, such as adjustment levels for actuators.

Input: The FNN receives real-time input from the sensor array, which is continuously monitoring the process variables:

- **Temperature sensors** provide data that are vital for preventing overheating and ensuring chemical processes proceed at optimal rates.
- **Pressure sensors** monitor the integrity of containment vessels and pipelines, preventing leaks and ruptures.

These inputs are sampled at a frequency high enough to allow real-time responses from the control system.

Results Interpretation: The outputs from the FNN directly influence the operational controls of the D1244. They are interpreted as follows:

- **Control actions:** Adjustments made by the control system based on FNN outputs are implemented immediately to optimize processes and reduce energy consumption.
- **Operational efficiency:** By continuously adapting to changing conditions, the D1244 maintains optimal performance with minimal waste of resources.
- **Safety:** The system enhances safety by maintaining all process variables within safe operational limits, reducing the risk of accidents.

All these features are summarized in a score value, which can be managed in the PHM solution for the prediction of future states.

4.3.4. Advantages and Future Prospects of Fuzzy Neural Networks

FNNs represent a significant advancement over traditional NNs in that they provide interpretable relationships despite the increased complexity of systems and data. This capability makes FNNs particularly valuable for modeling and controlling complex systems. The score generated by FNNs enables a simplified yet comprehensive representation of system states and interactions, which are summarized in a value that can be interpreted by the controller. This not only reduces complexity, but also lays the foundations for improved decision-making and system control.

The combination of the adaptability of fuzzy logic and the learning capacity of neural networks allows FNNs to process

imprecise and uncertain information effectively. This is particularly advantageous in decision-making processes where traditional methods reach their limits. FNNs show their strengths especially in scenarios characterized by high uncertainty and fuzziness and offer a robust solution to the challenges of the real world.

The future of FNNs in application methodology aims to achieve real-time data processing and interpretability. The further development of these technologies will enable even more precise decision-making, which will significantly improve the adaptivity and autonomy of the system components. This direction of development promises to increase efficiency and effectiveness in the processing of complex data structures and at the same time create the basis for innovative control and monitoring systems that are able to react dynamically to changes and make proactive decisions (Talpur et al., 2022).

The integration of FNNs with advanced data processing techniques, such as Long Short-Term Memory (LSTM), could further enhance the ability to analyze and adapt to newly added information in real time. This not only increases system performance, but also lays the foundation for extensive use of FNNs in a variety of applications, from predictive maintenance to optimization of operations (Wang, Shao, & Jumahong, 2023).

5. Conclusion

This paper provides a framework for the comprehensive realization of Prognostics and Health Management for a complex system of an adaptive high-rise building. This includes not only the load-bearing structure, but all domains involved in the system such as actuators, sensors, the control system as well as maintenance with all sub-domains such as spare parts stocking and maintenance resources.

Adaptive load-bearing structures offer the potential to increase the sustainability and service life of high-rise buildings. However, this integrates many other functional domains into the system that can fail and whose optimal use may require regular adjustments. Prognostics and Health Management offers the potential to reduce uncertainties about the current state of the system and to optimize it for use.

The PHM control loop consisting of five modeling elements – system, data, diagnosis, prognosis, and optimization – is used as the basis for the structure in the application of PHM. However, since embedding in a system is required for implementation, a framework is introduced here that does this under the requirements of the complex system of an adaptive structure.

The framework is divided into the areas of system analysis and modeling and the description of the PHM solution. The comprehensive nature of the framework and the systematic approach support the consideration and accurate integration of all functional areas of the adaptive structure into the PHM.

The system analysis consists of six sub-steps. First, the system is described, and the relevant system scopes are defined. On this basis, existing dimensions are identified. With the help of the dimensions and the system description, specifications, boundary conditions are determined. The target parameters for operation and its optimization are defined on the same basis. This is followed by the derivation of influencing factors on the target parameters and for compliance with boundary conditions. At the end of the system analysis, uncertainties are identified for compliance with the boundary conditions and optimization of the target parameters.

Modeling begins with the development of the target space and objective function. The target space describes all permissible solutions, as the boundary conditions are met. The objective function weights the individual target parameters in relation to each other if there are several target parameters. The next step is to develop the model structure. Here, the results of the system analysis are used again to map and link all the necessary functions. Finally, system integration is presented using the example of structural control for an adaptive high-rise building.

The PHM solution section covers details of the approaches for diagnosis, prognosis, and health management. A hybrid approach of physically based and data-driven models is recommended for diagnosis and prognosis to meet the requirements of the complex system. For the health management agent, the use of fuzzy neural networks is discussed to enable precise interpretation and processing of complex system states and dynamics.

Through the presented content, the paper provides, besides the framework for the implementation, the structure in terms of system sizes and models for the application of PHM to the complex system of the first adaptive high-rise building.

Acknowledgment

This work was supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft,
Project-ID 279064222), as a part of the collaborative research center CRC 1244 (SFB1244) “Adaptive Skins and Structures for the Built Environment of Tomorrow” Projects B03.

References

Dakova, S., Heidingsfeld, J. L., Böhm, M., & Sawodny, O. (2022). An optimal control strategy to distribute element wear for adaptive high-rise structures. In *2022 american control conference (acc)* (p. 4614-4619). doi: 10.23919/ACC53348.2022.9867396

de Campos Souza, P. V. (2020, July). Fuzzy neural networks and neuro-fuzzy networks: A review the

main techniques and applications used in the literature. *Applied Soft Computing*, *92*, 106275. doi: 10.1016/j.asoc.2020.106275

de Campos Souza, P. V., Lughofer, E., & Guimaraes, A. J. (2021, November). An interpretable evolving fuzzy neural network based on self-organized direction-aware data partitioning and fuzzy logic neurons. *Applied Soft Computing*, *112*, 107829. doi: 10.1016/j.asoc.2021.107829

Efinger, D., Ostertag, A., Dazer, M., Borschewski, D., Albrecht, S., & Bertsche, B. (2022). Reliability as a key driver for a sustainable design of adaptive load-bearing structures. *Sustainability*, *14*(2), 895.

Gienger, A., Schaut, S., Sawodny, O., & Tarin, C. (2020). Modular distributed fault diagnosis for adaptive structures using local models. *IFAC-PapersOnLine*, *53*(2), 13631–13637.

Goebel, K., Daigle, M., Saxena, A., Sankararaman, S., Roychoudhury, I., & Celaya, J. (2017). *Prognostics*. Wrocław: Amazon Fulfillment.

Henß, M. S. (2021). *Methodik zur konzeption, analyse und modellierung von lösungen im prognostics and health management (phm).* Universität Stuttgart.

Kim, N.-H., An, D., & Choi, J.-H. (2017). *Prognostics and health management of engineering systems*. Cham: Springer International Publishing.

Mishra, S., Sahoo, S., & Mishra, B. K. (2019). Neuro-fuzzy models and applications. In *Advances in computational intelligence and robotics* (p. 78–98). IGI Global. doi: 10.4018/978-1-5225-5793-7.ch004

OECD. (2015). *Material resources, productivity and the environment*. OECD Publishing.

Ostertag, A. (2021). *Zuverlässigkeit, sicherheit und nachhaltigkeit adaptiver tragwerke.* Universität Stuttgart.

Si, X.-S., Zhang, Z.-X., & Hu, C.-H. (2017). *Data-driven remaining useful life prognosis techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sobek, W., Haase, W., & Teuffel, P. (2000). Adaptive systeme. *Stahlbau*, *69*(7), 544–555.

Stiefelmaier, J., Böhm, M., Sawodny, O., & Tarín, C. (2023). Parity space-based fault diagnosis in piecewise linear systems*. *IFAC-PapersOnLine*, *56*(2), 10868-10873. (22nd IFAC World Congress) doi: https://doi.org/10.1016/j.ifacol.2023.10.764

Talpur, N., Abdulkadir, S. J., Alhussian, H., Hasan, M. H., Aziz, N., & Bamhdi, A. (2022, January). A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods. *Neural Computing and Applications*, *34*(3), 1837–1875. doi: 10.1007/s00521-021-06807-9

Thibaut Abergel, B. D., & Dulac, J. (2018). 2017 global status report for buildings and construction [Report]. *Report*. (Coordinated by United Nations Environment Programme. Supported by the governments of France and Germany, and the French Environment and Energy Management Agency (ADEME))

Wang, W., Shao, J., & Jumahong, H. (2023, November). Fuzzy inference-based lstm for long-term time series prediction. *Scientific Reports*, *13*(1). doi: 10.1038/s41598-023-47812-3

BIOGRAPHIES

Dshamil Efinger graduated as Master of Science in Mechanical Engineering at the University of Stuttgart, Germany, in 2019. Since then, he is pursuing his PhD studies as a Research Assistant in the Reliability Engineering Department at the Institute of Machine Components and the Technology Transfer Initiative at the University of Stuttgart. As a current member of the DFG-funded Collaborative Research Centre 1244 he investigates reliability, resilience, monitorability and safety of adaptive load-bearing structures. Find out more: linkedin.com/in/dshamil-efinger

Giuseppe Mannone graduated as Bachelor of Science in Mechanical Engineering at the University of Stuttgart, Germany, in 2022. Since 2023, he has been studying autonomous systems in the master's program at the University of Stuttgart, aiming for an M.Sc. He started working in reliability engineering at the Institute of Machine Components in 2020. Since 2022, he is focusing on the application of artificial intelligence in reliability engineering.

Martin Dazer, Dr.-Ing. received the B.Sc. degree in mechanical engineering from Baden-Württemberg Cooperative State University (DHBW), in 2011, the M.Sc. degree in mechanical engineering, in 2014, and the Dr.-Ing. degree in reliability engineering from the University of Stuttgart, in 2019. From 2015 to 2018, he was a Research Assistant at the Institute of Machine Components working on stochastic fatigue calculations and optimization methods in reliability test planning. He is currently the Head of the Reliability and Drive Technology Department at the Institute of Machine Components. He is also the Founder and a Consultant of RelTest-Solutions GmbH, Stuttgart, offering highly advanced reliability consulting, coaching, and training for industry. His current research interest includes the multitude of aspects of reliability engineering with its main focus on life testing. Dr. Dazer is a member of the Advisory Board of Safety and Reliability of the Association of German Engineers (VDI). He is also a member of the Program Committee of the Technical Reliability Conference, Germany. He is the Head of the Technical Committee of Reliability Management of VDI.

Automated Fault Diagnosis Using Maximal Overlap Discret Wavelet Packet Transform and Principal Components Analysis

Fawzi Gougam[1], Moncef Soualhi[2], Abdenour Soualhi[3], Adel Afia[4], Walid Touzout[1], and Mohamed Abdssamed Aitchikh[5]

[1] *LMSS, Faculté de technologie, Université de M'hamed Bougara Boumerdes, 35000 Boumerdes, Algeria*
f.gougam@univ-boumerdes.dz, w.touzout@univ-boumerdes.dz
[2] *Université de Franche-Comté, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France*
moncef.soualhi@univ-fcomte.fr
[3] *LASPI, IUT d Roanne, Université Jean Monnet university, 42300 Roanne, France*
abdenour.soualhi@univ-st-etienne.fr
[4] *Département de génie mécanique et productique, FGMGP, USTHB, 16111 Bab-Ezzouar, Algeria*
adel.afia@usthb.edu.dz
[5] *LEMI, Université de M'hamed Bougara Boumerdes, 35000 Boumerdes, Algeria*
ma.aitchikh@univ-boumerdes.dz

ABSTRACT

Bearings and gears are components most susceptible to failure in electromechanical systems, especially rotating machines. Therefore, fault detection becomes a crucial step, as well as fault diagnosis. Over decades, substantial progress in this field has been observed and numerous methods are now proposed for feature extraction from monitoring data. Among these data, vibration signals are most used. However, in the presence of non-Gaussian noise, most conventional methods may be inefficient. In this paper, a hybrid methodology is proposed to address this potential issue. The proposed methodology uses a combination of the Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) and Principal Component Analysis (PCA) techniques. First, the MODWPT technique decomposes the vibration signal with uniform frequency bandwidth, facilitating effective signal processing and introducing diversity for enhanced time-frequency signals. Then, to identify significant patterns and characteristics related to faults, PCA is used for 3D dimensional representation of system health state by capturing the variance in the extracted features. Subsequently, a self-organizing map (SOM) is used for system state classification for diagnostics. This technique is applied to open-access test bench data containing vibration signals with non-Gaussian noise.

Keywords: Signal processing, Fault diagnosis, Gearbox, Feature extraction, Rotating machines, MODWPT.

1. INTRODUCTION

Bearings and gears have significant roles in both traditional and modern manufacturing processes due to their extensive applicability (A. Soualhi et al., 2014; Benaggoune et al., 2020). As technology progresses rapidly, bearings and gears are often used in industrial equipment and systems (Gougam et al., 2021). Consequently, any bearing or gear failure can have a profound impact on the entire production process, resulting in economic losses and potentially fatal accidents. Consequently, early bearing and gear failure diagnosis becomes a critical step in preventing premature and catastrophic failures throughout the manufacturing process. Detecting and addressing gear failures is fundamental in preventing serious economic losses and potential accidents. A robust failure detection and isolation technique is required to monitor the rotational components and identify any damage (Abdeltwab & Ghazaly, 2022). Currently, many researchers are focusing on monitoring conditions using vibration signals. Various conventional methods, such as the Fast Fourier Transform, the Wigner Ville distribution (Dhok et al., 2020), short-time Fourier transform (S. Zhou et al., 2020), and cyclo-stationary analysis (Gilles, 2013; Adel et al., 2022; Patel & Upadhyay, 2020), have been employed. However, background noise often obscures defect-characteristic information, resulting in non-stationary, non-linear behavior of the data signal. Hence, such methods are regarded as ineffective in extracting features indicating early defects. Several adaptive decomposition methods have been introduced in recent decades for feature extraction. As a promising example, empirical mode decomposition (EMD), proposed by Huang, has been widely explored and applied to mechanical fault diagnosis (Meng et al., 2022). EMD decomposes a discrete-time signal into in-

dividual components called intrinsic mode functions (IMFs). Each IMF includes an oscillating component with several frequency levels, and the determination of the sifting iteration number is determined by a sifting stop criterion (SSC) (Fan et al., 2020). In the decomposition process, EMD uses a specialized filter with a bandwidth and center frequency that adjust dynamically according to the signal's characteristics.

Nevertheless, HHT presents two major drawbacks derived from EMD: mode mixing, in which waves of the same frequency are assigned to different intrinsic mode functions (IMFs), and end effects, which lead to incorrect instantaneous values at both ends of the signal (Afia, Gougam, Rahmoune, et al., 2023). A more appropriate decomposition method needs to be considered. Consequently, wavelet analysis has returned to the center of attention, in particular the discrete wavelet transform (DWT), widely used in condition monitoring and fault diagnosis to extract time-frequency features (Syed & Muralidharan, 2022). Unfortunately, DWT requires the sample size to be precisely 2 (down-sampling) during the analysis (Wang et al., 2021). An enhanced version of the discrete wavelet transform, known as the Maximal Overlap Discrete Wavelet Transform (MODWT), handles the sampling reduction process, yet still remains plagued by inadequate frequency resolution, similar to that of the discrete wavelet transform [14]. To address such limitations, the Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) has replaced both MODWT and DWT, providing improved resolution. MODWPT decomposes the complex signal into single components of instantaneous amplitude and frequency, ensuring circular shift equivariance to monitor the gear's working condition (Afia et al., 2024a). For automated health monitoring, various machine learning techniques are used to provide more accurate predictions (M. Soualhi et al., 2021; Lourari et al., 2024; Benaggoune et al., 2022; M. Soualhi et al., 2019; Gougam, Afia, Aitchikh, et al., 2024; M. Soualhi et al., 2020; Afia, Gougam, Touzout, et al., 2023; Tahi et al., 2020; Touzout et al., 2020; A. Soualhi et al., 2012; Gougam, Afia, Soualhi, et al., 2024; Afia et al., 2024b). The self-organizing map (SOM) belongs to the artificial neural network (ANN) category, trained by unsupervised learning. SOM objective is to generate a low-dimensional - typically two-dimensional - discretized representation called a map, offering a dimensionality reduction method in the input space of training samples (Z. Zhou et al., 2020; Zhang et al., 2020). In this context, SOM offers a significant advantage, as it improves data interpretability. By clustering in grids and reducing dimensionality, data become more accessible, facilitating the identification of underlying patterns (Fan et al., 2021). In this paper, MODWPT was used as the signal processing method for decomposition. Afterwards, PCA was applied to reduce the dimensions while preserving the stable signal features. A comparison with EMD-PCA is presented to assess the advantages of the proposed approach. The final step in the proposed methodology incorporates the use of the self-organizing map (SOM) for defect clustering. SOM, a neural network-based algorithm, categorizes faults on the basis of extracted features, providing a sophisticated clustering process. This enhances defect detection and analysis by offering a nuanced exploration of relationships and patterns in vibration data.

2. PROPOSED METHODOLOGY

This section aims to present the different steps of the proposed methodology. First, raw data are injected to the Maximal Overlap Discret Wavelet Packet Transform (MODWPT) to extract features by filtering the signal and enhance their dimensionality. Then, the obtained new data passes through Principal Component Analysis (PCA) for reducing dimensionality by converting intricate vibration signals into a collection of uncorrelated principal components. After that, a 3D representation of the three principal components (PCs) of different health state under variable working conditions will be generated. Finally, a self-organizing map (SOM) is used to classify the different patterns for faults diagnosis. And overall view of the proposed technique is presented in Figure 1.

3. MAXIMAL OVERLAP DISCRET WAVELET PACKET TRANSFORM

The raw data is initially segmented into 25 groups, each consisting a length of 10024 samples. This segmentation is carried out as a preliminary step for data augmentation, a process aimed at enhancing the dataset's diversity and robustness by generating additional instances. The MODWPT uses segmented raw data as input for multi-stage filtering, resulting in a greater number of vibration time-frequency bands.Similar to Mallat's algorithm (Afia et al., 2024a), MODWPT relies on quadrature mirror filters. The filters, denoted as and, represent a low-pass and a high-pass filter, each with a length of L=10024 samples (assumed to be even). For this purpose, 16 resulting wavelet coefficients (decomposed signals) are obtained from MODWPT filters, as presented in Equation 1.

$$\left\{\begin{array}{l} \sum_{l=0}^{L-1} \tilde{h}_l^2 = \frac{1}{2} \\ \sum_{l=0}^{L-1-2k} \tilde{h}_l \tilde{h}_{l+2k} = 0, \quad k = 1, 2, ..., \frac{1}{2}(L-2) \\ \tilde{g}_l = (-1)^{l+1} \tilde{h}_{L-l-1}, \quad l = 0, 1, ..., L-1 \end{array}\right\} \quad (1)$$

MODWPT diverges from Mallat's method by employing interpolation two-based decimation operation. More precisely, at each MODWPT level, $(2^{j-1} - 1)$ zeros are introduced between two consecutive adjacent coefficients of $\widetilde{g}_l$ and $\widetilde{h}_l$. This ensures that the wavelet coefficients generated (WT) for each wavelet sub-band maintain an equivalent length to that

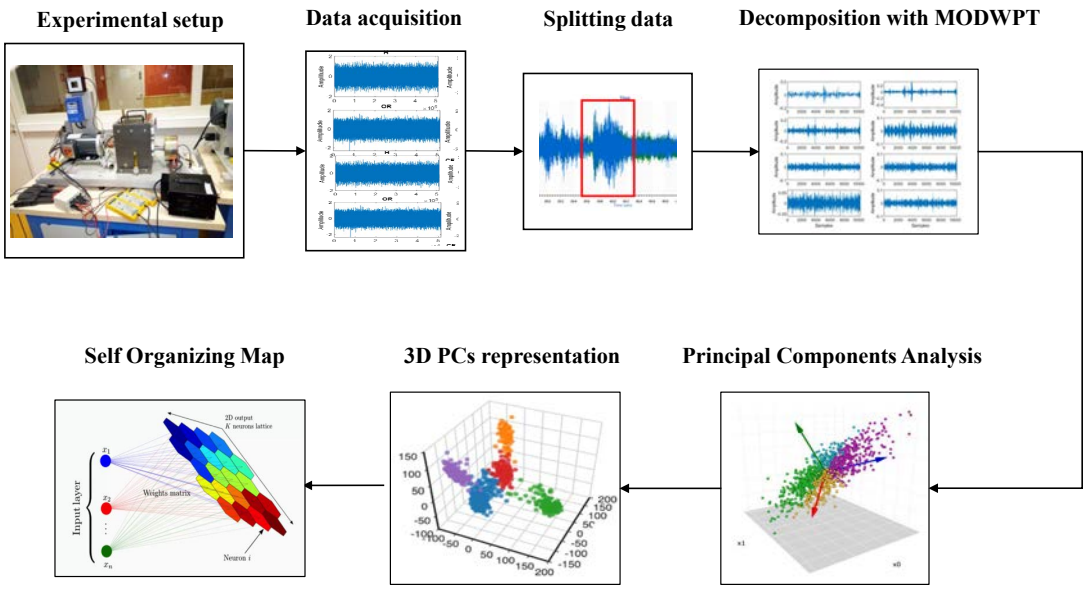


Figure 1. Overall view of the proposed technique.

of the input signal (Afia et al., 2024a). Considering a discrete-time sequence $x(t), t = 0, 1, \ldots\ldots, , N-1$, where N represents the sequence length, the wavelet coefficients $W_{j,n,t}$ of the nth sub-band at level j are computed using the following equations, with n taking values from $0 to 2^j - 1$. The initial condition is given by $W_{0,0,t} = x(t)$ [14]. For a discrete-time sequence$x(t), t = 0, 1, \ldots\ldots\ldots, , N-1$, where N is the sequence length, the wavelet coefficients$W_{j,n,t}$ of the nth sub-band at level j are calculated according to the following equations in which $n = 0, 1, ..., 2^j - 1, W_{0,0}, t = x(t),\ t = 0, 1, ..., N-1$ (Afia et al., 2024a):

$$W_{j,n,t} = \sum_{l=0}^{L-1} \tilde{f}_{n,l}\ W_{j-1,[n/2](t-2^{j-1}l) \bmod N} \tag{2}$$

$$\tilde{f}_{n,l} = \begin{cases} \tilde{g}_l, & \text{if } n \bmod 4 = 0 \text{ or } 3 \\ \tilde{h}_l, & \text{if } n \bmod 4 = 1 \text{ or } 2 \end{cases} \tag{3}$$

4. Principal Components Analysis

This step of the methodology aims to exploit the wavelets coefficients data for anomaly detection. For this purpose, the obtained data matrix, comprising 16 wavelet coefficients with 10024 samples(16,10024), is fed into PCA for dimensionality reduction and 3D visualization. In fact, conventional literature works aims to use extracted features and directly train Machine Learning (ML) models for classification. This procedure lacks efficiency with regard to train a ML model, already considered as a black box, with no verified and reliable feature. In such a scenario, PCA generates stable Principal Components from various vibration health state signals. PCA is proving advantageous in vibration signal analysis for fault diagnosis due to its multiple benefits (Shi et al., 2020). PCA excels in dimensionality reduction, transforming complex vibration signals into a set of uncorrelated principal components. This simplifies subsequent analysis, improves computational efficiency and provides a concise representation of essential information. PCA contributes to noise reduction in vibration signals. By emphasizing the principal components associated with the greatest variances, PCA actually mitigates the noise impact, making it particularly useful in environments where the signal-to-noise ratio is a significant concern. PCA can be conceptualized as an unsupervised learning problem. The process of deriving principal components from a raw dataset can be simplified into six steps:

1. **B**egin with the entire dataset, initially comprising d+1 dimensions, and disregard the labels, resulting in a new dataset of d dimensions.
2. **C**alculate the mean for each dimension across the entire dataset.
3. **C**ompute the covariance matrix for the complete dataset; where i is the samples number of signal X. are the mean of X,Y signals.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1} \tag{4}$$

4. **D**etermine the eigenvectors and their corresponding eigenvalues.

$$Cov(X,Y) \times \sum\nolimits_{Value} = \sum\nolimits_{Value} \times \sum\nolimits_{Vector} \tag{5}$$

5. **A**rrange the eigenvectors in descending order based on eigenvalues, selecting k eigenvectors with the highest

eigenvalues to create a d×k dimensional matrix, denoted as W.

6. Utilize this d × k eigenvector matrix (W) to transform the samples to obtain the new set of uncorrelated variables.

5. Self Organizing Map

A Self-Organizing Map (SOM) is an unsupervised machine learning algorithm used for clustering and visualization of high-dimensional data. SOM is employed to identify patterns and anomalies in complex systems. SOMs consist of a grid of nodes (neurons) organized in two dimensions [29]. Each neuron is associated with a weight vector that is adjusted during the learning process. The best matching unit (BMU) weights and its neighboring neurons are adjusted to move closer to the input pattern. Neighboring neurons in the SOM grid respond similarly to similar input patterns, forming clusters (Figure 2).

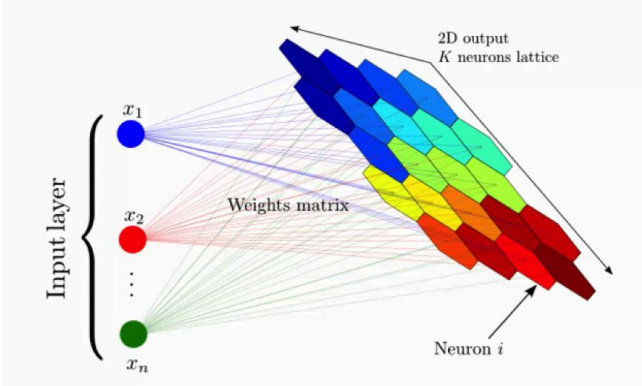


Figure 2. Self-organizing map architecture.

6. Application and Results

The described methodology is applied to experimental data, which includes various fault states as well as a healthy state. The experimental setup is designed for multi- faults classification. With the proposed methodology, our objective is to evaluate the effectiveness of the extracted features in separating the different health states.

6.1. Case Study

To verify the applicability of the proposed methodology, an open access data of test bench provided and presented in (M. Soualhi et al., 2023). The test bench chosen in this study is Laboratoire d'Analyse des Signaux et Processus Industriels (LASPI) benchmark that introduce bearing and gear fault detection and diagnostic problem(Figure 3). It consists of a three-phase inverter controlling a 1.5 kW induction motor driving the gearbox. An electromagnetic brake connected to the gearbox simulates the motor load.

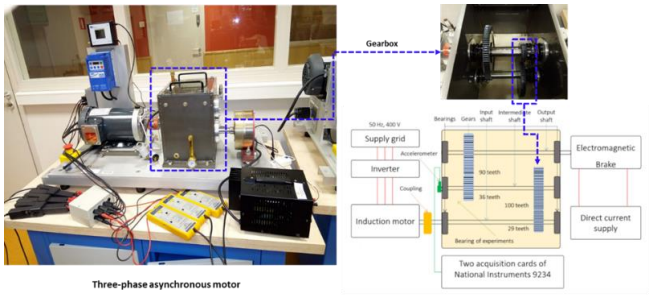


Figure 3. LASPI Benchmark.

The gearbox consists of three shafts: input, intermediate and output, with the studied bearings located on the intermediate shaft. The input shaft, connected directly to the motor, features a 29-tooth gear and two bearings with 9 balls each. The bearings have a 0.3125" diameter, a 1.5157" pitch diameter and a 0" contact angle.

Measurements are conducted in continuous mode for 10 seconds, with a 25.6 kHz sampling frequency. Vibration signals are acquired with an accelerometer sensor with a 100 mV/g sensitivity. Using the specified test rig instrumentation, a total of five distinct health states were examined, encompassing healthy bearings, inner, outer ring defects and combined bearing defects. In addition, each condition was tested at two different speeds: 25 Hz. Furthermore, each speed condition was tested at two load levels: 0%, 50% and 75%.

6.2. Result and discussion

The used raw data represent different states of bearing and gears. Each signal is first splitted into 25 segments. Figure 4 displays the different acquired vibration signals. Subsequent

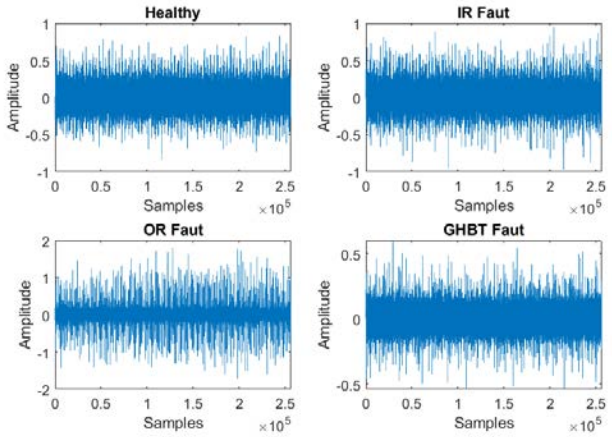


Figure 4. Vibration signals of different health states.

steps involve applying MODWPT to the segments in order to decompose each signal into its different frequency components, thus obtaining a detailed signal representation in the

time and frequency domains(Figure 5).

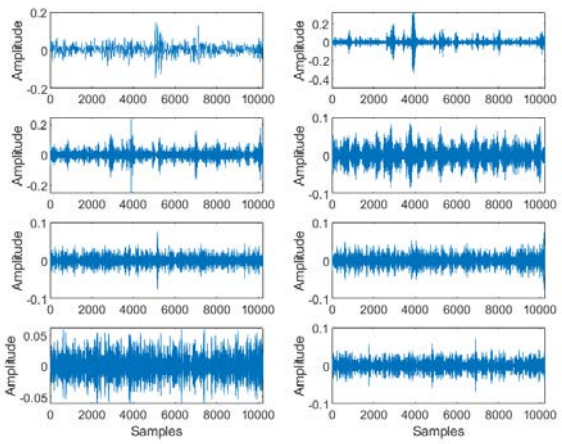


Figure 5. MODWPT decomposition.

PCA is applied to the decomposed modes, enabling a more concise data display by focusing on the dominant features. This step is crucial for simplifying the data set while retaining the essential information, thus facilitating analysis and interpretation.

Figure 6 illustrates the sample distribution under variable working conditions $(0\%, 50\% and 75\% load)$ using EMD-PCA. In the visual representation depicted in Figure 6, a noticeable level of sample confusion is obvious. This intricacy involves complex patterns and overlapping samples in the dataset, leading to a significant degree of imprecision in the extraction process. Recognizing the critical nature of this issue, our proposed solution is a hybrid MODWPT-PCA technique.

Figure 7 provides a more accurate sample distribution between the different health states. In comparison with the EMD-PCA, the advantages of the proposed approach in the feature extraction step are affirmed, demonstrating the proposed methodology's effectiveness. For automated fault diagnosis, a self-organizing map (SOM) is used to cluster neighboring neurons that respond similarly to analogous input patterns. The extracted data from MODWPT-PCA which containing 100 samples (25 samples for each 4 health states) and three columns (3 Principal Components) is used as input data for SOM model. Figure8 shows the sample clustering map using the proposed methodology for the four health states under three distinct working conditions.

After examining the map clustering (Figure 8), a significant sample separation is seen across distinct health states under different working conditions. This validates the reliability of the suggested diagnostic methodology.

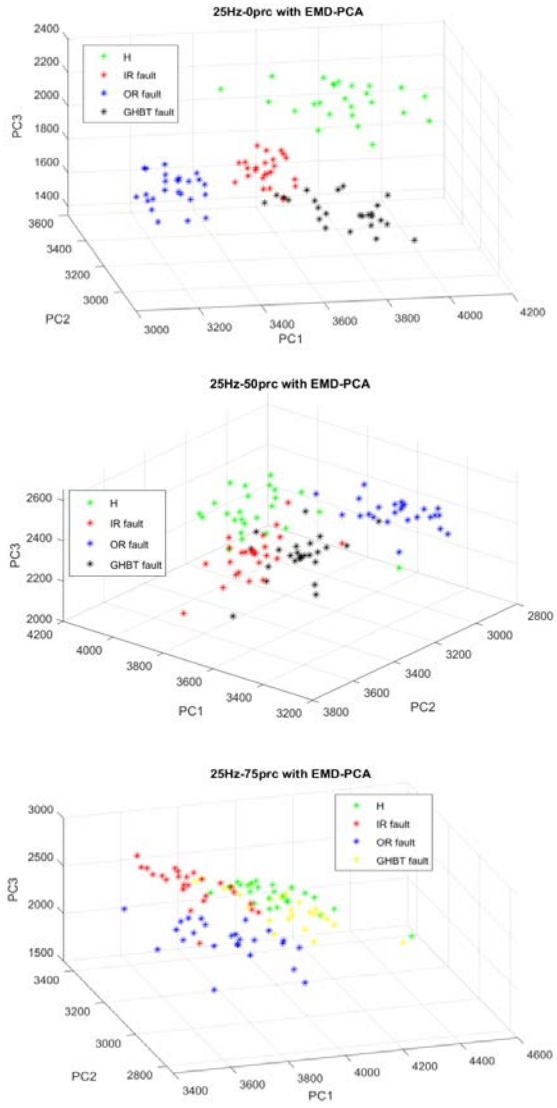


Figure 6. Samples distribution using EMD-PCA.

7. Conclusion

In this paper, the authors proposed a hybrid technique to address the inherent limitations of hidden fault characteristics in fault diagnosis. The approach integrates Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) and Principal Component Analysis (PCA). MODWPT efficiently decomposes data signals with uniform frequency bandwidth, while PCA proves advantageous for feature extraction in vibration signal analysis. PCA captures variance, enabling the identification of significant defect-related patterns. A comparison with EMD-PCA is then conducted to assess the performance of the suggested algorithm. Finally, a self-organizing map (SOM) is used for machine learning to cluster the acquired data samples. The experimental results highlight that the proposed methodology is highly efficient in extract-

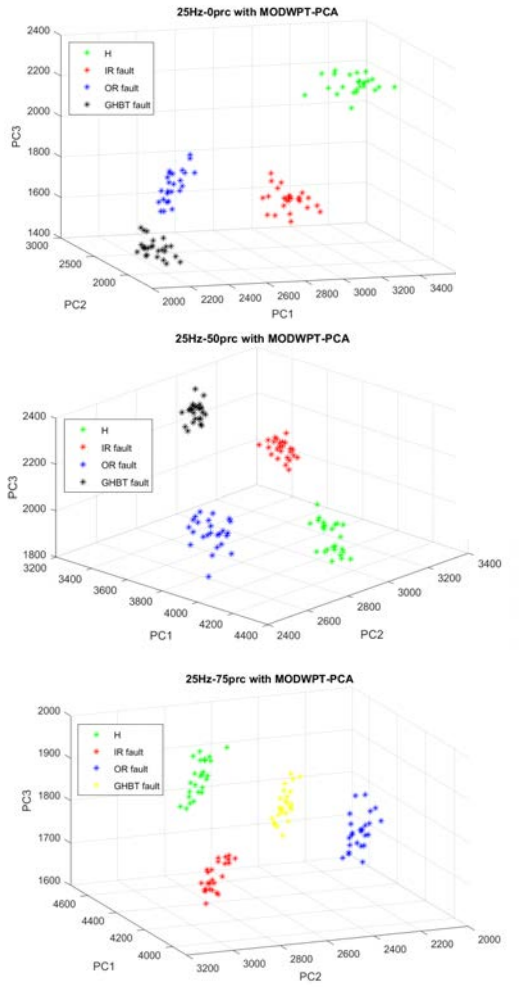


Figure 7. Samples distribution using MODWPT-PCA.

ing fault signatures from raw vibration data, even in a complex working environment.

REFERENCES

Abdeltwab, M. M., & Ghazaly, N. M. (2022). A review on engine fault diagnosis through vibration analysis. *International Journal on Recent Technologies in Mechanical and Electrical Engineering*, *9*(2), 01–06.

Adel, A., Hand, O., Fawzi, G., Walid, T., Chemseddine, R., & Djamel, B. (2022). Gear fault detection, identification and classification using mlp neural network. In *Recent advances in structural health monitoring and engineering structures: Select proceedings of shm and es 2022* (pp. 221–234). Springer.

Afia, A., Gougam, F., Rahmoune, C., Touzout, W., Ouelmokhtar, H., & Benazzouz, D. (2023). Gearbox fault diagnosis using remd, eo and machine learning classifiers. *Journal of Vibration Engineering & Technologies*, 1–25.

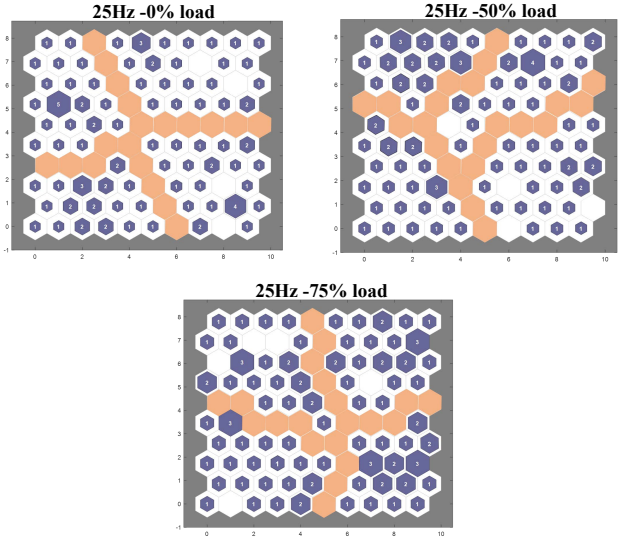


Figure 8. SOM clusters map.

Afia, A., Gougam, F., Rahmoune, C., Touzout, W., Ouelmokhtar, H., & Benazzouz, D. (2024a). Intelligent fault classification of air compressors using harris hawks optimization and machine learning algorithms. *Transactions of the Institute of Measurement and Control*, *46*(2), 359–378.

Afia, A., Gougam, F., Rahmoune, C., Touzout, W., Ouelmokhtar, H., & Benazzouz, D. (2024b). Intelligent fault classification of air compressors using harris hawks optimization and machine learning algorithms. *Transactions of the Institute of Measurement and Control*, *46*(2), 359–378.

Afia, A., Gougam, F., Touzout, W., Rahmoune, C., Ouelmokhtar, H., & Benazzouz, D. (2023). Spectral proper orthogonal decomposition and machine learning algorithms for bearing fault diagnosis. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *45*(10), 550.

Benaggoune, K., Meraghni, S., Ma, J., Mouss, L., & Zerhouni, N. (2020). Post prognostic decision for predictive maintenance planning with remaining useful life uncertainty. In *2020 prognostics and health management conference (phm-besançon)* (pp. 194–199).

Benaggoune, K., Yue, M., Jemei, S., & Zerhouni, N. (2022). A data-driven method for multi-step-ahead prediction and long-term prognostics of proton exchange membrane fuel cell. *Applied Energy*, *313*, 118835.

Dhok, S., Pimpalkhute, V., Chandurkar, A., Bhurane, A. A., Sharma, M., & Acharya, U. R. (2020). Automated phase classification in cyclic alternating patterns in sleep stages using wigner–ville distribution based features. *Computers in Biology and Medicine*, *119*, 103691.

Fan, H., Shao, S., Zhang, X., Wan, X., Cao, X., & Ma, H. (2020). Intelligent fault diagnosis of rolling bearing using fcm clustering of emd-pwvd vibration images. *IEEE Access*, *8*, 145194–145206.

Fan, H., Yan, Y., Zhang, X., Cao, X., & Ma, J. (2021). Composite fault diagnosis of rolling bearing based on optimized wavelet packet ar spectrum energy entropy combined with adaptive no velocity term pso-som-bpnn. *Journal of Sensors*, *2021*, 1–15.

Gilles, J. (2013). Empirical wavelet transform. *IEEE transactions on signal processing*, *61*(16), 3999–4010.

Gougam, F., Afia, A., Aitchikh, M., Touzout, W., Rahmoune, C., & Benazzouz, D. (2024). Computer numerical control machine tool wear monitoring through a data-driven approach. *Advances in Mechanical Engineering*, *16*(2), 16878132241229314.

Gougam, F., Afia, A., Soualhi, A., Touzout, W., Rahmoune, C., & Benazzouz, D. (2024). Bearing faults classification using a new approach of signal processing combined with machine learning algorithms. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *46*(2), 65.

Gougam, F., Chemseddine, R., Benazzouz, D., Benaggoune, K., & Zerhouni, N. (2021). Fault prognostics of rolling element bearing based on feature extraction and supervised machine learning: Application to shaft wind turbine gearbox using vibration signal. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *235*(20), 5186–5197.

Lourari, A. w., Soualhi, A., Medjaher, K., & Benkedjouh, T. (2024). New health indicators for the monitoring of bearing failures under variable loads. *Structural Health Monitoring*, 14759217231219486.

Meng, D., Wang, H., Yang, S., Lv, Z., Hu, Z., & Wang, Z. (2022). Fault analysis of wind power rolling bearing based on emd feature extraction. *CMES-Computer Modeling in Engineering & Sciences*, *130*(1), 543–558.

Patel, S. P., & Upadhyay, S. H. (2020). Euclidean distance based feature ranking and subset selection for bearing fault diagnosis. *Expert Systems with Applications*, *154*, 113400.

Shi, H., Guo, J., Yuan, Z., Liu, Z., Hou, M., & Sun, J. (2020). Incipient fault detection of rolling element bearings based on deep emd-pca algorithm. *Shock and Vibration*, *2020*, 1–17.

Soualhi, A., Clerc, G., & Razik, H. (2012). Detection and diagnosis of faults in induction motor using an improved artificial ant clustering technique. *IEEE Transactions on Industrial Electronics*, *60*(9), 4053–4062.

Soualhi, A., Medjaher, K., & Zerhouni, N. (2014). Bearing health monitoring based on hilbert–huang transform, support vector machine, and regression. *IEEE Transactions on instrumentation and measurement*, *64*(1), 52–62.

Soualhi, M., El Koujok, M., Nguyen, K. T., Medjaher, K., Ragab, A., Ghezzaz, H., ... Ouali, M.-S. (2021). Adaptive prognostics in a controlled energy conversion process based on long-and short-term predictors. *Applied Energy*, *283*, 116049.

Soualhi, M., Nguyen, K. T., & Medjaher, K. (2020). Pattern recognition method of fault diagnostics based on a new health indicator for smart manufacturing. *Mechanical Systems and Signal Processing*, *142*, 106680.

Soualhi, M., Nguyen, K. T., Soualhi, A., Medjaher, K., & Hemsas, K. E. (2019). Health monitoring of bearing and gear faults by using a new health indicator extracted from current signals. *Measurement*, *141*, 37–51.

Soualhi, M., Soualhi, A., Nguyen, K. T., Medjaher, K., Clerc, G., & Razik, H. (2023). Open heterogeneous data for condition monitoring of multi faults in rotating machines used in different operating conditions. *International Journal of Prognostics and Health Management*, *14*(2).

Syed, S. H., & Muralidharan, V. (2022). Feature extraction using discrete wavelet transform for fault classification of planetary gearbox–a comparative study. *Applied Acoustics*, *188*, 108572.

Tahi, M., Miloudi, A., Dron, J., & Bouzouane, B. (2020). Decision tree and feature selection by using genetic wrapper for fault diagnosis of rotating machinery. *Australian Journal of Mechanical Engineering*.

Touzout, W., Benazzouz, D., Gougam, F., Afia, A., & Rahmoune, C. (2020). Hybridization of time synchronous averaging, singular value decomposition, and adaptive neuro fuzzy inference system for multi-fault bearing diagnosis. *Advances in Mechanical Engineering*, *12*(12), 1687814020980569.

Wang, M.-H., Lu, S.-D., & Liao, R.-M. (2021). Fault diagnosis for power cables based on convolutional neural network with chaotic system and discrete wavelet transform. *IEEE Transactions on Power Delivery*, *37*(1), 582–590.

Zhang, J., Wu, J., Hu, B., & Tang, J. (2020). Intelligent fault diagnosis of rolling bearings using variational mode decomposition and self-organizing feature map. *Journal of Vibration and Control*, *26*(21-22), 1886–1897.

Zhou, S., Xiao, M., Bartos, P., Filip, M., & Geng, G. (2020). Remaining useful life prediction and fault diagnosis of rolling bearings based on short-time fourier transform and convolutional neural network. *Shock and Vibration*, *2020*, 1–14.

Zhou, Z., Chen, J., Zi, Y., & An, T. (2020). A modified som method based on nonlinear neural weight updating for bearing fault identification in variable speed condition. *Journal of Mechanical Science and Technology*, *34*, 1901–1912.

Bayesian Networks for Remaining Useful Life Prediction

Erik Hostens, Kerem Eryilmaz, Merijn Vangilbergen, and Ted Ooijevaar

Flanders Make, Gaston Geenslaan 8, 3001 Heverlee, Belgium
erik.hostens@flandersmake.be
kerem.eryilmaz@flandersmake.be
merijn.vangilbergen@flandersmake.be
ted.ooijevaar@flandersmake.be

ABSTRACT

Remaining useful life (RUL) prediction is a critical task in the field of condition-based maintenance. It is important to perform RUL prediction in a statistical sound way. However, it is not straightforward to properly combine multiple information sources about an asset, such as available statistics, measurements, derived features, and prior knowledge in the form of mathematical models and relations, including their uncertainties. Bayesian networks (BNs) are a means of graphically representing all statistical information in a comprehensible way and allow for correctly combining all information. BNs allow for inference in all directions, thereby not merely providing a RUL prediction with explicit uncertainty, but select the most informative features, diagnose which degradation mechanism is manifest if multiple mechanisms exist, provide decision support in the form of optimal condition-based maintenance points when combined with a cost model. BNs also explicitly quantify the model uncertainty arising from the scarcity of the training data. We illustrate these benefits on two real-world industrial examples: solenoids and bearings. We also provide a method to correctly include the effect of changing operating conditions.

1. INTRODUCTION

Condition-based maintenance (CBM) has gained a strong interest from the industry in recent years, both driven by the market-driven necessity of ever-increasing efficiency and sustainability of industrial systems, and the opportunity opened up by the fast growing industrial digitization and sensorization. For an extensive recent literature survey, we refer to (Quatrini, Costantino, Di Gravio, & Patriarca, 2020). The prediction of remaining useful life (RUL), which is the time at which an industrial asset will have been degraded to such extent that it can no longer perform its intended function, plays an important role in CBM to schedule maintenance, optimize operating efficiency, and above all avoid unplanned downtime.

Given the intrinsic randomness of the drivers of degradation leading up to the ultimate failure of the asset, a proper statistical treatment of these phenomena is required (Sankararaman, 2015). Indeed, because of the costly consequence of an unanticipated failure, even when its probability is small, the expected cost may become significant and requires early action. Many performance metrics for RUL prediction focus on the deviation of mean prediction from the ground truth, but the decision support for maintenance actions should rather focus on the tail of the prediction distribution.

In this paper, we aim to present a generic and systematic method using *Bayesian networks* (BNs) to incorporate all available knowledge and data in the RUL prediction. When many factors contribute to this prediction, it is a challenge to manage these relations and correctly calculate the overall statistics. BNs, although in essence nothing more than a representation of the statistics, offer a comprehensible approach. The explicit modeling and quantification of RUL prediction uncertainty has been extensively studied in literature, typically focused on a particular industrial asset, such as (Mishra, Martinsson, Rantatalo, & Goebel, 2018) for batteries and (Prakash, Narasimhan, & Pandey, 2019) for bearings.

A Bayesian network is a graphical representation of a joint distribution of a set of variables (Pearl, 1988). The joint distribution is factorized into root probabilities and conditional dependencies, which are graphically represented by a directed acyclic graph. BNs are ideal for taking an event that occurred and assessing the probability that any one of several possible known causes was the contributing factor. For example, a BN could represent the probabilistic relationships between failure mechanisms and their manifestations in the sensor data. Given sensor data, the network can be used to compute the probabilities of the presence of various failure mechanisms. This is analogous to how medical doctors need

to diagnose a patient showing symptoms of disease, see for example Fig. 1. In summary, BNs are a means for graphically representing complex statistical relations that become otherwise intractable.

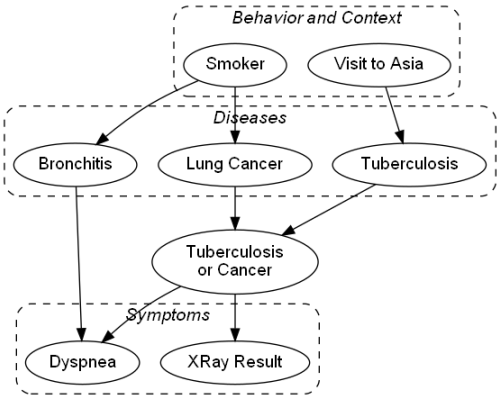


Figure 1. An example of a BN in medicine, analogous to condition-based maintenance. The intuitive (causal) understanding of how stochastic variables relate is well captured: diseases (failure mechanisms) are caused by behavior and context (settings and operating conditions) and become manifest in observed symptoms (measurements or features derived thereof). In this particular example, smoking increases the probability of bronchitis or lung cancer, whereas if a patient recently visited Asia, there is a higher probability of having contracted tuberculosis. The symptoms alone cannot distinguish tuberculosis from cancer, therefore an auxiliary variable "Tuberculosis or Cancer" is used to make this explicit.

The purpose of this paper is to show the advantages of using BNs for RUL prediction: (i) BNs naturally combine all prior knowledge, in the form of models and statistics, and data, and as such maximally exploit the available information; (ii) distributions of unknown variables can be inferred from observed variables, in all possible directions, depending on the application (parameter estimation, diagnostics, prognostics, decision support); (iii) explicit model uncertainty, which can be used to assess the (in)sufficiency of the available training data; (iv) BNs are easily extended with more variables that affect RUL and its assessment, such as operating conditions. We will illustrate how BNs are used on two examples of assets widely used in industry, a solenoid-operated valve (SOV) and a bearing, on which we conducted accelerated life testing.

This paper is organized as follows. In Section 2, we introduce the two industrial assets, SOVs and bearings, on which we validated the BN method for RUL prediction. In Section 3, we explain the methodology of BNs for RUL prediction and show how they provide the aforementioned advantages. In particular, we explain (i) how the model is built, trained and used, (ii) how a cost model can be integrated in the BN to provide *decision support for maintenance*, (iii) how *model uncertainty* is taken into account in the RUL prediction, and (iv) how the BN is adjusted to incorporate *operating conditions*. In Section 4, we illustrate the methodology by the application on the two industrial assets in four case studies, corresponding to the topics (i)-(iv). Finally, we formulate some conclusions and future directions of research in Section 5.

2. The application cases

2.1. Solenoid-operated valves

We have conducted our research on a historic dataset of accelerated life tests (ALT) on a set of 3/2-way normally closed alternating current powered solenoid operated valves (SOV). The SOVs were subjected to on-off switch cycles until failure or end-of-life (EOL), defined as the moment that the solenoid's magnetic force is insufficient to overcome the friction and move the plunger. This moment is observed both in the current signal, as the solenoid then behaves as a fixed nonlinear inductance, and in the thermal mass flow detecting leakage measured at the outlet ports and blow-off holes of the valves. The experimental dataset has been used before in (Tod et al., 2019; Mazaev, Ompusunggu, Tod, Crevecoeur, & Van Hoecke, 2020), where full details can be found.

In previous work, we have defined a number of features on the current signal and quantitatively assessed their feature performance for health monitoring quantitatively. For a detailed description of these features, we refer to (Ompusunggu & Hostens, 2021, 2023). For our purposes, it suffices to know that some features can be extracted from the current signal that contain information on the state of health of the SOV. Without loss of generality, we do not distinguish between direct measurement or derived features, we call them both *measurements*. Fig. 2 shows an important measurement *time-to-hit*, defined as the time between start of induced current in the solenoid at the beginning of the cycle and the plunger hitting the end of the shaft stopping its movement. In the induced increasing current signal, this stopping of the plunger is seen as a small dip in the first period. The time-to-hit increases when the SOV degrades, as it is related to the friction between plunger and shaft and therefore indicative of health. Fig. 3 shows the evolution of time-to-hit for 10 SOVs as a function of the number of past on-off cycles, together with the corresponding EOLs, except for two solenoids that did not fail before the end of the ALT. Note the strong increase of time-to-hit approaching the EOL.

2.2. Bearings

The SOVs were all tested under the same operating conditions. In order to illustrate the BN methodology for RUL prediction under *varying operating conditions*, we reused datasets of ALTs that we have conducted on bearings (Geurts, Eryilmaz, & Ooijevaar, 2023). These were generated using the Flanders Make Smart Maintenance living lab, an open test and development platform that aims to support the adoption

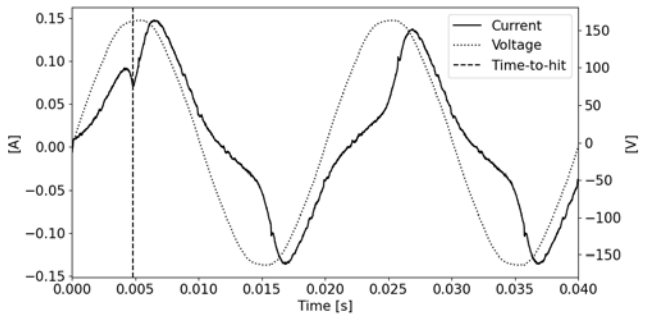


Figure 2. Definition of time-to-hit measurement in the SOV's alternating current signal.

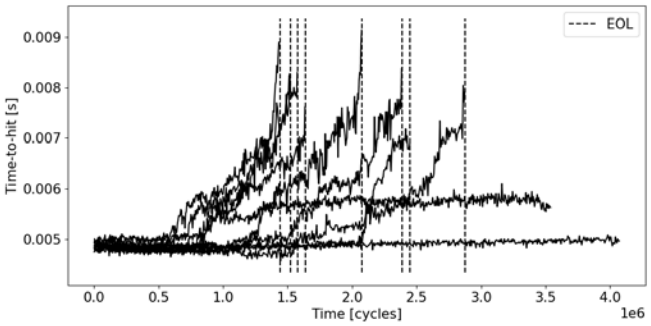


Figure 3. Evolution of time-to-hit for 10 SOVs. Two solenoids did not reach their EOL during the ALT.

of condition monitoring technologies in industry (Ooijevaar, Di, et al., 2019). Details of the setup and the tests can be found in the given references, we restrict ourselves here to the information required for a self-contained comprehension of this paper.

Before the start of each ALT, a small initial indentation was created in the bearing inner race in a repeatable manner. This serves as a local stress riser emulating a local plastic deformation caused by, for instance, a contamination particle. The EOL of a bearing is defined as the moment where the measured vibrations exceed a peak-to-peak acceleration of about 200 m/s^2. One set of bearings was subjected to stationary operating conditions, being a radial load of 9 kN and a rotary speed of 2000 rpm, another set of bearings to the same radial load but a varying speed going from 1000 to 2000 rpm in a cyclic saw-tooth pattern with a period of about 10 minutes, as shown in Fig. 4. Note how the acceleration depends on the speed, and how it increases exponentially with time near the EOL, similarly as the time-to-hit for the SOVs.

3. METHODOLOGY

3.1. Building the BN

The BN defines the joint distribution of all considered random variables X_i as a product of the individual density functions, conditional on their parent variables $\mathrm{pa}(i)$, i.e. the variables

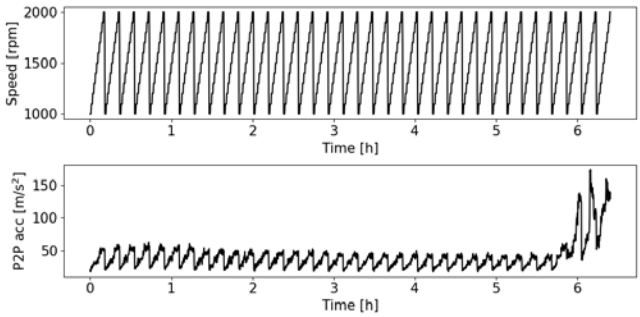


Figure 4. Speed and peak-to-peak acceleration of a bearing subjected to varying rotary speed.

that point to variable X_i in the graph representation:

$$p(X) = \prod_i p\left(X_i | X_{\mathrm{pa}(i)}\right). \quad (1)$$

We show how BNs naturally combine all available information for RUL prediction. The information we consider comprises:

- lifetime statistics,
- a degradation model,
- measurements revealing the underlying level of degradation.

A lot of research has been spent to each of these elements of information, either generic or specific to the considered asset. It is the sole purpose of this paper to show how these are *combined*, so we make a few simple assumptions, that sufficiently fit our example.

- In the following, we will refer to *time* not in a literal sense, but rather expressed in a unit that naturally relates to the usage of an asset. For solenoids, it is the number of on-off cycles; for bearings, it is the number of rotations. Similarly, lifetime and RUL are expressed in the same unit.
- For lifetime statistics, we assume a Weibull distribution, that corresponds to a failure rate that is proportional to a power of time (Jiang & Murthy, 2011). On top of that, in cases where the asset has not yet failed, we know that lifetime is greater than the current time.
- As a degradation model, we assume a hidden dimensionless degradation state, where the rate of degradation is proportional to the level of degradation itself. This simple first order dynamics boils down to an exponentially increasing degradation, or equivalently an exponentially decreasing health, which intuitively corresponds to the well-known P-F curve (Nowlan & Heap, 1978). The relation between health and measurements is a function that we will preferably describe with only a few parameters so as to keep complexity low, but the method allows any function fitting algorithm, including neural net-

works. Here we will adopt simple linear relations with normally distributed random noise, motivated by the exponential decay of health and the measurement evolution plots shown in Figs. 3-4.

These relations are captured in the generic BN structure of Fig. 5. Some of the variables are shown in rectangles, to indicate that they are deterministic, either always observed or a deterministic function of their parent variables. This simple model is captured in the following explicit relations:

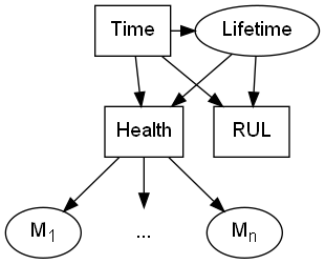


Figure 5. The generic BN structure for RUL prediction: time is given, and is a lower bound for the lifetime. Their difference is the RUL, and also defines the hidden health state. Measurements $M_1, \ldots, M_n$ are function of health.

$$\begin{aligned} L &\sim \text{Weibull}\,(k, \lambda)\,, \\ L &> T, \\ \text{RUL} &= L - T, \\ H &= 1 - \exp\left(\frac{T-L}{D}\right), \\ M_i &\sim \mathcal{N}\left(M_{i0}(1-H) + M_{i1}H, \sigma_i^2\right), \end{aligned} \tag{2}$$

where L denotes the lifetime, T time, and H health. The latter starts very close to 1 (at $T = 0$) and ends on 0 (at EOL, or $T = L$). This BN model leverages on expert knowledge captured in simple relations between the variables and is therefore capable of describing those relations using only a few parameters, as opposed to using e.g. neural network models that easily have hundreds of free parameters. The free parameters are in this case: the Weibull distribution shape k and scale λ, the degradation time D, which approximately corresponds to the time between onset of degradation and end-of-life, and for each measurement M_i the spread σ_i and the linear coefficients M_{i0}, M_{i1}, which are the mean measurement values at start and at EOL, respectively. One could wonder why health H is made a deterministic variable, and not a random variable. We motivate this by the fact that by defining lifetime L and the measurements M_i as random variables, all *real* stochasticity is already captured. Indeed, as H is never observed directly, it can be considered merely as an auxiliary variable linking the M_i with L and T. Its actual value is of no importance, unless it would have an impact on the (observed) performance, but here we only consider the EOL and the prediction thereof.

3.2. Training and prediction

The purpose of *training* the BN model is to fit the parameter values to the data from the ALTs. We can then use this model to *predict* RUL for new data of another asset. To this end, we have used PyMC, a probabilistic programming library for Python that allows users to build Bayesian models and fit them using *Markov chain Monte Carlo (MCMC)* methods (Patil, Huard, & Fonnesbeck, 2010). Essential to PyMC is that there is no distinction between parameters and variables, there are only (random) variables, including the parameters. This enables *Bayesian hierarchical modeling*, a type of Bayesian modeling where information is available on different levels (Allenby & Rossi, 2006). In our case, we assume a single set of parameters k, λ, M_{i0}, M_{i1}, D, σ_i for the entire population, but we have a different L for each asset, and H, RUL, M_i are different for each asset and each time T.

MCMC is used for both training and prediction, they only differ in which variables are observed and which not. This is graphically explained in Fig. 6, showing all variables including the parameters (but only one measurement, to not overload the picture), and marking their being observed as gray shading. In training, the parameters are unknown and are fitted to the data on lifetime and measurements. In prediction, the parameters are known, but the unknown lifetime distribution is to be inferred from the lifetime statistics and the measurements.

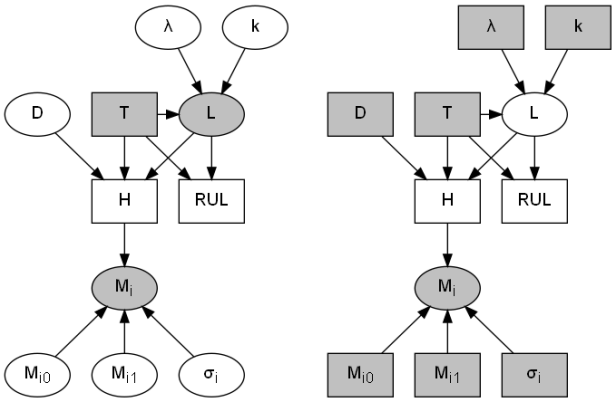


Figure 6. The BNs for training (left) and prediction (right). Gray shading represents known/observed variables, whereas white represents unknown/unobserved variables.

The MCMC algorithm updates the distributions of all unobserved random variables given their prior distributions and the data, through the likelihood of the data. The prior distributions, also called prior *beliefs*, quantify the uncertainty in the prior knowledge *before* data is acquired. It is based on previous data campaigns or *physical* and *statistical* knowledge of the asset's behavior and degradation. For instance, the shape parameter k is related to the trend of the failure rate, typically going up ($k > 1$) as the asset ages. If little is

known beforehand, prior beliefs should be chosen sufficiently wide, so-called *weakly informed priors*. The updating of prior beliefs into posterior distributions when new data comes in is the central paradigm of Bayesian statistics. Therefore, it is important to note that the BN for prediction in Fig. 6 is an oversimplification: the parameters do not become exactly known by the training, but if their posterior distributions become sufficiently narrow, their low remaining uncertainties can be ignored in their contribution to the total uncertainty of the RUL prediction.

Another important nuance to make about Fig. 6 is the fact that, as we saw in Section 2, some of the assets' EOL is never observed, simply because their ALT is stopped early. So not all L-nodes in the left graph (training) of Fig. 6 should be gray shaded (observed). Such *censoring* is quite common in statistical analysis of survival data in medicine (Kalbfleisch & Prentice, 2011). Because some of the lifetimes L are not observed, their values cannot be directly used to infer the Weibull parameters. This is exemplary for why the BN framework is powerful: although L is not observed, the uncertainty on its unknown value can nonetheless be significantly reduced through its relation with the measurements and the time during which it did not fail. Therefore this information still contributes to the fitting of the Weibull parameters. The BN truly leverages on all the available information combined.

3.3. Decision support for maintenance scheduling

We explain how the BN for prediction, shown in Fig. 6, is adjusted to provide decision support for maintenance scheduling. We only consider asset replacement as the maintenance action, but an analogous reasoning can be followed for other maintenance actions. The BN prediction yields a probability distribution of RUL. This is more useful than a single RUL expected value, because it allows to better balance the risk of unanticipated failure with the economic loss of early replacement. The BN allows the integration of a cost model and evaluate the prediction of cost given the replacement scheduling strategy. We illustrate this with a simple cost model. We assume the asset's cost C_A in the normal situation. This is for instance the sum of the costs of purchase and installation, the latter coinciding with the scheduled replacement of its used predecessor. If the asset fails before its scheduled replacement, there will be an extra *cost of failure* C_F. This cost is very dependent on the application: it can be very high for high impact failures, such as significant production loss or damage to other equipment, but it can also be low or even zero. In that case, the asset should only be replaced after EOL.

Another parameter we assume in this example is the *prediction horizon* T_{PH}, defined as the time required to schedule the replacement up front, for instance because it takes some time to send a maintenance engineer to the asset's remote location. If it is decided at time T to replace the asset, the actual replacement can take place at time $T+\Delta T$, where $\Delta T \geq T_{\mathrm{PH}}$. This is schematically depicted in Fig. 7.

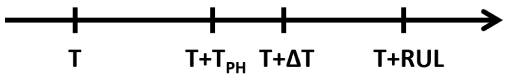


Figure 7. Replacement is scheduled at time T, and takes place at time $T+\Delta T$, which is $T+T_{\mathrm{PH}}$ at the earliest. $T+\Delta T$ should precede $T+\mathrm{RUL}$ to avoid the cost of failure.

Let $T+\Delta T$ be the asset's *replacement time* in the future, decided at the *scheduling time* T hence depending on the RUL prediction at time T. The resulting total cost depends on the actual RUL:

$$\begin{aligned} \text{total cost} &= C_A && \text{if RUL} \geq \Delta T, \text{ or} \\ &= C_A + C_F && \text{if RUL} < \Delta T. \end{aligned} \tag{3}$$

To balance the extra cost C_F with the cost of early replacement, where more assets are used in the long run, we have to evaluate the *cost per used time unit* C_T:

$$\begin{aligned} C_T &= \frac{C_A}{T+\Delta T} && \text{if RUL} \geq \Delta T, \text{ or} \\ &= \frac{C_A + C_F}{T+\mathrm{RUL}} && \text{if RUL} < \Delta T. \end{aligned} \tag{4}$$

Note that C_T is a deterministic function of other variables. We include it in the BN for prediction, shown in Fig. 8. Alongside the RUL prediction at scheduling time T, we can use this model to calculate the distribution of the cost $C_T(\Delta T)$ corresponding to the replacement time $T+\Delta T$, for multiple values of ΔT. Replacement should then be scheduled as soon as the

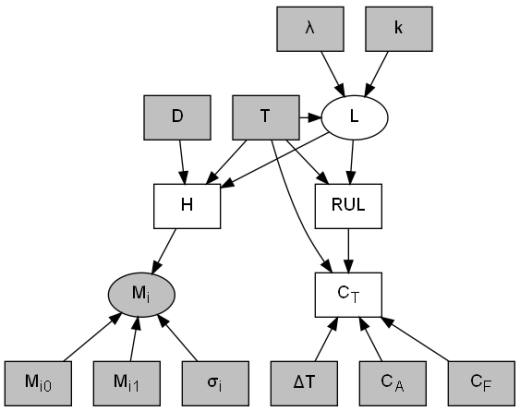


Figure 8. The BN for prediction, including the simple cost model.

expected value of this cost $E[C_T(\Delta T)]$ reaches a minimum for $\Delta T = T_{\mathrm{PH}}$, or if:

$$E[C_T(\Delta T)] > E[C_T(T_{\mathrm{PH}})], \quad \forall \Delta T : \Delta T > T_{\mathrm{PH}}. \tag{5}$$

3.4. Model uncertainty

As we mentioned in Section 3.2, the training of the BN will result in a posterior distribution of all non-observed random variables, including the model parameters. If these parameter posteriors are sufficiently narrow, we may consider the remaining uncertainty insignificant and select the means of the posteriors as fixed known parameters for the prediction BN, as was shown in Fig. 6. However, if training data is scarce, the remaining posterior uncertainty cannot be ignored and should be included in the prediction. This is achieved by making the parameter nodes in the prediction BN stochastic and unobserved, and using the posterior distribution after training as its prior. It is important to note that, in many cases, the parameter distributions will be mutually *dependent* after training. Therefore, one should use a single joint prior distribution for the parameters in the prediction BN.

MCMC does not output explicit posterior distributions, but a sample thereof, due to the way it works. To include it in the prediction, there are two options: either combine training and prediction in one MCMC run, or approximate the posterior parameter distribution. The former option is the most correct, since in this way we are combining all information at once, both of the past ALTs and the running one. However, this requires a lot of calculations and all data need to be kept, so this approach may become cumbersome. The latter option is most practical, since there is only one run of MCMC involving the training data, after which they are not longer needed. For approximating the posterior joint distribution of the parameters, in most cases a multivariate normal distribution is suited, motivated by the fact that, if the model is well designed, it is expected that the parameter estimates will converge. To this end, PyMC also supports *automatic differentiation variational inference (ADVI)* as an alternative to MCMC followed by the approximation of the posterior distribution from the sample. ADVI turns this around by up front assuming a parameterized approximation of the posterior distribution and reformulating its calculation as an optimization problem (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2017).

3.5. Varying operating conditions

In our original BN model of Section 3.1, we assumed the asset's degradation as the sole driver of further degradation. This works fine if other influences do not have a significant contribution to degradation. However, in most cases, the *operating conditions* (OC) do have a strong impact on degradation, and should be taken into account. Secondly, the OC also influence the measurements. This is clearly seen in Fig. 4 for the bearings. This influence further complicates the analysis, as the measurements serve as indicators for degradation, so we need to distinguish whether changes in the measurement are resulting either by changed OC, or by degrading health, or both.

Inquiring the effect of OC on degradation is particularly difficult, since the OC consist of multiple variables that often have a combined effect where one OC variable strengthens or weakens the effect of another. In such cases, on the one hand a detailed understanding is needed of how the asset's health evolves under given OC, in the form of engineering laws or physical models. On the other hand, a sufficient amount of ALT data is required to validate and quantify these models. However, ALT data are typically scarcely available because they are costly to generate. Again, maximally leveraging on all available knowledge and data is key. We show how the original BN for RUL prediction is adjusted such as to account for OC.

Let us first address the simpler case of *stationary* operating conditions. When the OC are stationary over the entire lifetime, even when they are different for different assets, training and using a BN for prediction of RUL is not a lot more complicated than before. Given sufficient ALT data for each OC, one can simply retrain another BN for each different OC. Of course it is more useful to leverage on knowledge of how OC affect lifetime, ideally in the form of a relation with the OC adding few free parameters, such as the empirical basic rating life model for bearings of (ISO281, 2007). Such knowledge is integrated in the BN by adding a relation from the OC to the lifetime and the measurements. The generic BN for RUL prediction, previously shown in Fig. 5, is then updated to the BN structure of Fig. 9.

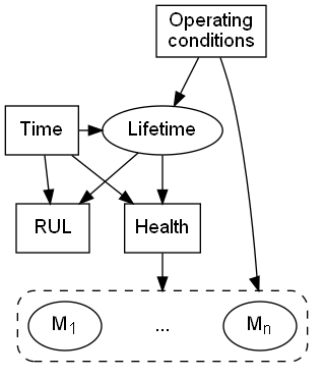


Figure 9. The generic BN structure for RUL prediction under stationary operating conditions.

For *varying* operating conditions, the relation between OC and lifetime is more complicated. One could interpret the relation between OC and lifetime in Fig. 9 as the *aggregated effect* of OC on lifetime, but as such the combined effect of the OC with health is overlooked. For instance, a higher load might have a larger damaging effect if the asset was already in a degraded state. Therefore, the OC effect should be aggregated in such a way that it takes that aspect into account, which it does not in the BN structure of Fig. 9.

We resolve this issue by assuming *nominal* parameters for a single nominal OC defined up front. We relate all other OC to this nominal OC, and express lifetime and RUL as their equivalent lifetime and RUL under the nominal OC. To deal with varying OC, we locally *compress* and *stretch* time into an equivalent time under the nominal OC. This idea is shown in the BN structure of Fig. 10. The variables Health, Lifetime and RUL are all expressed in the *Equivalent Time* corresponding to nominal OC. For the nominal OC, Equivalent Time and Time progress at the same rate. Note that Equivalent Time is defined as a stochastic variable (ellipse), because the relation between OC, Time and Equivalent Time might be uncertain. This BN structure also includes the immediate effect of OC on the measurements. The RUL that this model predicts is expressed as the equivalent RUL under nominal OC. Decision support should take the *expected future* OC into account for predicting the actual RUL. Note that such BN can also be used to recommend to change the future OC, if the application allows it, in order to delay potential failure.

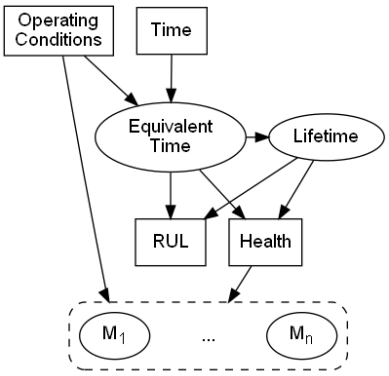


Figure 10. The generic BN structure for RUL prediction under varying operating conditions.

4. Application on case studies

In this section, we illustrate the methodology by the application on either SOVs or bearings in four case studies, corresponding to the topics explained in the previous section: (i) training and prediction, (ii) decision support for maintenance, (iii) model uncertainty, and (iv) varying operating conditions.

4.1. BNs for RUL prediction on SOV

We have trained the BN model of Fig. 6 on the ALT data of the 10 SOVs, whose time-to-hit measurement evolution was shown in Fig. 3, with 2 ALTs censored. For simplicity, we have only incorporated the time-to-hit measurement. Including other measurements would only reduce the prediction uncertainty, although not significantly since time-to-hit is the most informative on the hidden health. The BN automatically weighs the measurement contributions according to the amount of information they provide on the health. As such, it is an implicit form of *feature extraction*. We then used the fitted model parameters to predict the RUL of an SOV not used in the training, over its entire lifetime. The prediction yields a distribution of RUL, the evolution of which is shown in Fig. 11, compared to the RUL ground truth. Note the sud-

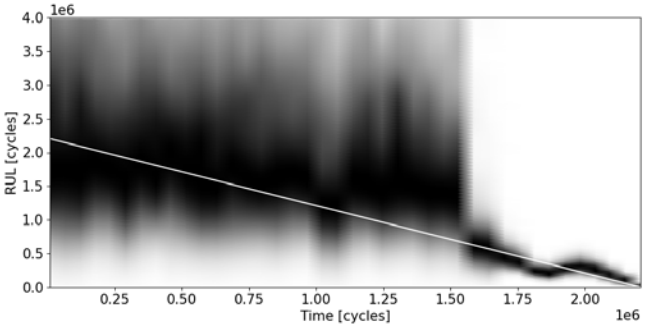


Figure 11. RUL prediction for one SOV (probability density in grayscale) compared to the ground truth RUL (white line).

den decrease of uncertainty around 1.5 million cycles. This decrease is due to the fact that at that point, the time-to-hit measurement starts increasing, thus providing crucial information on the imminent EOL. Before, the measurement reveals little on the SOV's health, so the prediction is mainly based on the Weibull statistics, truncated at the current cycle. This transition is naturally taken care of by the BN because it combines all information sources available and automatically weighs their uncertainties in the statistical posterior, as opposed to an explicit switching such as the one used in (Geurts et al., 2023).

The RUL prediction shown in Fig. 11 is based on the last measured time-to-hit at the present cycle. However, the time-to-hit measurement itself displays stochastic fluctuations, as can be clearly seen in Fig. 3. Therefore, it makes sense to include the *full history* of the measurement in the RUL prediction, so that this inherent stochasticity is filtered. Yet it is important to note that the fluctuations are not white noise, rather colored noise, which means the measurements are correlated over time. M_i should then no longer be defined as separate univariate normal random variables like in Eq. 2, but as a single multivariate normal random vector with the same mean and the measurement autocovariance as covariance. The resulting prediction ($5\% - 95\%$ quantiles near the EOL) is shown in Fig. 12, for both the last measurement only and the full history prediction, illustrating the advantage of the latter: it is more consistent and accurate compared to the true RUL.

4.2. BNs for maintenance decision support on SOV

For the SOV example of Fig. 11, we show a detail of the RUL prediction approaching EOL and the corresponding expected relative cost $E[C_T(T_{\mathrm{PH}})]/C_A$ in Fig. 13, for arbitrary cost model parameters $C_F/C_A = 10$ and $T_{\mathrm{PH}} = 2e4$. Note that

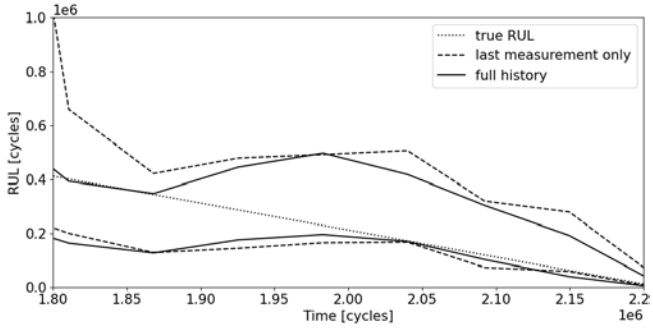


Figure 12. $5\%-95\%$ quantiles for RUL predictions from only the last measurement or from the full measurement history.

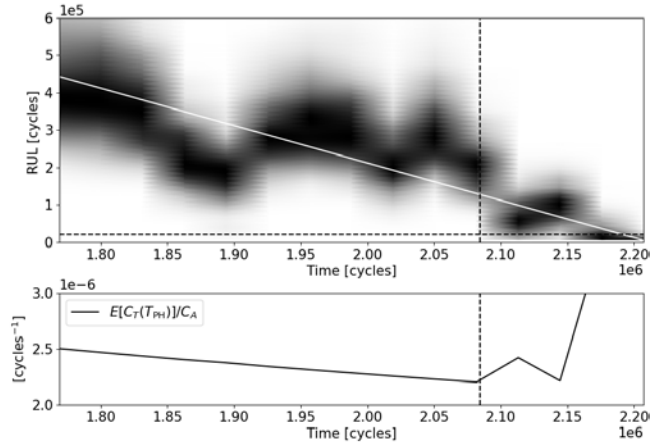


Figure 13. Top: RUL prediction. The dashed horizontal line marks the prediction horizon T_{PH}. Bottom: expected relative cost of replacement time $T + T_{\mathrm{PH}}$ at scheduling time T. The dashed vertical line marks the optimal scheduling time T^*.

at the optimal scheduling time T^*, the RUL prediction distribution still has the most part *above* the T_{PH} line. The optimal maintenance scheduling strategy therefore involves probing the *tail* of the RUL prediction distribution, which emphasizes the importance of correctly calculating this distribution.

4.3. Model uncertainty in BNs on SOV

We have redone the prediction of the SOV of Fig. 11, now using 40 SOVs in the training set instead of only 10. Both are compared in Fig. 14 through their $5\% - 95\%$ quantiles. In the RUL predictions, we have now included the posterior parameter uncertainty after training. Clearly, more training data results in a more accurate RUL prediction. It can be seen that this effect is most manifest in the healthy phase of the SOV, where the RUL prediction is mainly based on the EOL statistics and not on the measurement. This is to be expected from a statistical perspective. However, approaching the EOL, where the prediction accuracy is more important for optimal maintenance scheduling, both predictions become very close. This illustrates the power of BNs for RUL prediction, as in this case it suffices to have a training set of only 10 ALTs, two of which are censored, and a simple degradation and measurement model.

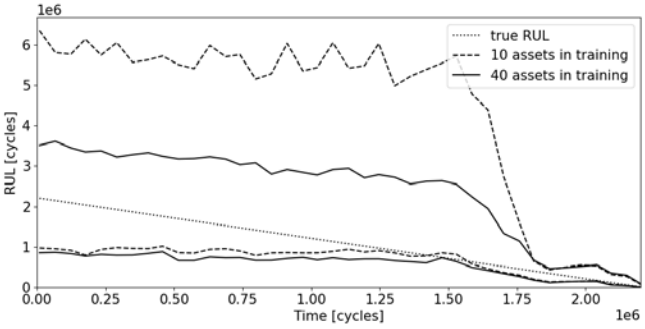


Figure 14. $5\%-95\%$ quantiles for RUL predictions including parameter uncertainty, where the BN parameters are trained either on 10, or on 40 SOVs.

4.4. BNs for varying OC on bearings

The bearing ALT dataset introduced in Section 2.2 is insufficiently rich to validate the proposed BN for RUL prediction under varying OC of Fig. 10. Indeed, because of the very uniform saw-tooth pattern of speed (Fig. 4) in the varying speed ALT, its long-term influence on degradation effectively corresponds to a stationary OC, albeit different from the stationary speed ALT. As a consequence, we have only two different long-term aggregated OC. We therefore use a combination of the BN structures of Figs. 9-10 for stationary OC and varying OC, respectively: we assume the long-term effect of OC on lifetime as equivalent to stationary, but the immediate effect of OC on the P2P measurement as varying. The resulting BN for training, now including the parameters,

is shown in Fig. 15. There are two extra variables SET and

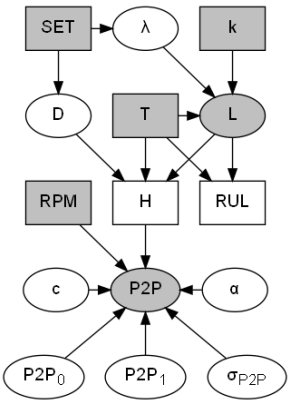


Figure 15. The BN for training of bearing RUL prediction, either under stationary or under varying RPM.

RPM, that relate to the aggregated effect of speed on lifetime and to the immediate effect of speed on the P2P acceleration measurement, respectively. The binary variable SET defines to which dataset the bearing belongs, either subjected to stationary speed ALT or to varying speed ALT. This variable essentially selects either one of two values for the parameters D and λ. The shape parameter k was fixed up front to a value of 1.3 building on historical knowledge on bearing fatigue lifetime statistics (NSWC, 2011). Extending Eq. 2 defining the relations between all variables, we define the distribution of the P2P measurement as:

$$\mathrm{P2P}_{\mathrm{nom}} \sim \mathcal{N}\left(\mathrm{P2P}_0(1-H)+\mathrm{P2P}_1 H, \sigma^2_{\mathrm{P2P}}\right), \quad (6)$$

$$\mathrm{P2P} = c\,\mathrm{P2P}_{\mathrm{nom}}\,\mathrm{RPM}^{\alpha}. \quad (7)$$

The expression in Eq. 7 with parameters c and α was established through a qualitative inspection of P2P data, both in healthy and degrading state as shown in Fig. 4, by comparing P2P values to the corresponding nominal P2P values around the nearest time where the speed is 2000 rpm. All 7 parameters of the BN model are simultaneously fitted to the training data.

We have trained this BN model on a set of 48 ALT, of which 7 were subjected to the varying speed profile. Bearings to validate the resulting RUL prediction were left out of the training data. As a benchmark, we also trained and validated the original model of Fig. 6 on the same data. An example of the resulting RUL prediction for both models on the same varying speed bearing is shown in Fig. 16. The RUL prediction with the original model is clearly disturbed by the varying conditions, emphasizing the need for including them. The same problem was manifest in the work of (Geurts et al., 2023).

A single asset's prediction may illustrate the added value, yet a proper comparison should be built on adequate RUL prediction performance metrics. A thorough overview and analysis of metrics is given in (Saxena, Celaya, Saha, Saha, & Goebel, 2010). To keep things simple, we have compared the benchmark model and the varying OC model by the log-likelihood evaluated at various relative locations in the lifetime and averaged over the varying speed bearings, as shown in Fig. 17. The log-likelihood is a straightforward generalization of mean squared error (MSE) that also covers the uncertainty of the prediction. The evaluation at multiple relative locations in the lifetime, similar to the alpha-lambda performance metric, addresses the application-specific prediction horizon as explained in Section 3.3.

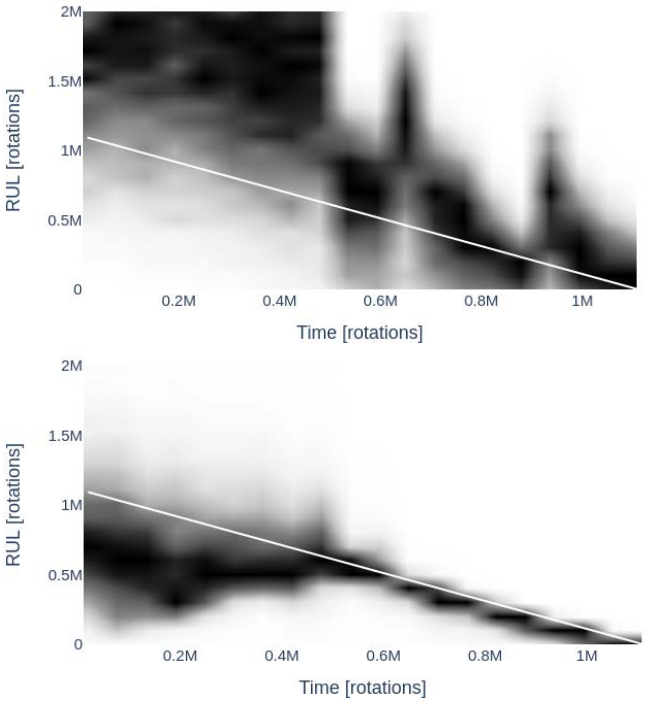


Figure 16. The RUL prediction for a varying speed bearing, both for the benchmark model (top) and the new model that includes the effect of varying OC (bottom), compared to the true RUL (white line).

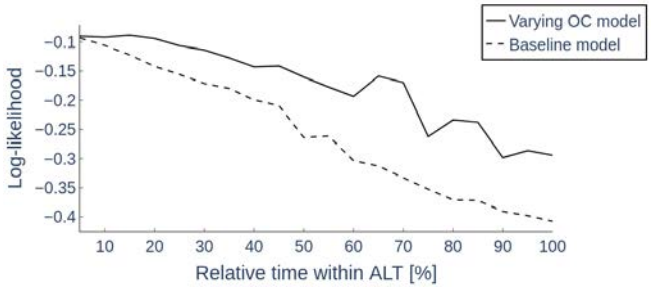


Figure 17. The average log-likelihood for the benchmark model and the new model, as function of relative time within the ALT.

5. CONCLUSION

We have shown a method for building, training and using Bayesian networks for RUL prediction. Next to the advantage of its comprehensibility, even when many factors contribute to the prediction, we have focused on the extension of RUL prediction with decision support for maintenance and the explicit inclusion of model uncertainty arising from the scarcity of training data. We have shown how the BN is adjusted to allow for RUL prediction under varying operating conditions.

This work is part of a larger study on the application of BNs for CBM and for maintaining quality in industry. We see the following open challenges and future research topics:

- Our current ALT datasets on SOVs and bearings do not allow for a proper validation of the generic method of Section 3.5. To this end, we are currently conducting a new ALT data campaign on SOVs under varying operation conditions.
- The inclusion of model uncertainty and its propagation to the RUL prediction is still lacking a quantified decision support for further data campaigns and design-of-experiments (DoE). We will investigate a practical method to assess the need for more training data and DoE, for instance through a criterion on the trend of a suitable performance metric such as the average leave-one-out log-probability.
- Investigate more complex degradation mechanisms, arising from multiple root causes that have different degradation dynamics.
- Instead of focusing on RUL which assumes the asset's quality as a binary variable and the EOL as a specific moment in time, we will shift towards the prognostics of a more nuanced application-oriented quality condition and corresponding decision support, such as condition-aware control.

ACKNOWLEDGMENT

This work has been carried out within the framework of Flanders Make's Strategic Basic Research project QUASIMO (Quality via a System Intelligence Methodology). Flanders Make is the Flemish strategic research centre for the manufacturing industry.

REFERENCES

Allenby, G. M., & Rossi, P. E. (2006). Hierarchical bayes models. *The handbook of marketing research: Uses, misuses, and future advances*, 418–440.

Geurts, K., Eryilmaz, K., & Ooijevaar, T. (2023). A sequential hybrid method for full lifetime remaining useful life prediction of bearings in rotating machinery. In *Annual conference of the phm society* (Vol. 15).

ISO281. (2007). *Rolling bearings-dynamic load ratings and rating life.* ISO: International Organization for Standardization.

Jiang, R., & Murthy, D. (2011). A study of weibull shape parameter: Properties and significance. *Reliability Engineering & System Safety*, *96*(12), 1619–1626.

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*, *18*(14), 1–45.

Mazaev, G., Ompusunggu, A. P., Tod, G., Crevecoeur, G., & Van Hoecke, S. (2020). Data-driven prognostics of alternating current solenoid valves. In *2020 prognostics and health management conference (phm-besançon)* (pp. 109–115).

Mishra, M., Martinsson, J., Rantatalo, M., & Goebel, K. (2018). Bayesian hierarchical model-based prognostics for lithium-ion batteries. *Reliability Engineering & System Safety*, *172*, 25–35.

Nowlan, F. S., & Heap, H. F. (1978). Reliability-centered maintenance.

NSWC. (2011). *Handbook of reliability prediction procedures for mechanical equipment*. Naval Surface Warfare Center West Bethesda, MD.

Ompusunggu, A. P., & Hostens, E. (2021). Physics-inspired feature engineering for condition monitoring of alternating current-powered solenoid-operated valves. In *International conference on maintenance, condition monitoring and diagnostics* (pp. 139–151).

Ompusunggu, A. P., & Hostens, E. (2023). Quantitative evaluation of electric features for health monitoring and assessment of ac-powered solenoid operated valves. *IFAC-PapersOnLine*, *56*(2), 3725–3731.

Ooijevaar, P. K., Ted and, Di, Y., et al. (2019). Smart machine maintenance enabled by a condition monitoring living lab. In *8th ifac symposium on mechatronic systems (mechatronics 2019) and the 11th ifac symposium on nonlinear control systems (nolcos 2019)* (Vol. 52). Elsevier.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, *35*(4), 1.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Prakash, G., Narasimhan, S., & Pandey, M. D. (2019). A probabilistic approach to remaining useful life prediction of rolling element bearings. *Structural health monitoring*, *18*(2), 466–485.

Quatrini, E., Costantino, F., Di Gravio, G., & Patriarca, R. (2020). Condition-based maintenance—an extensive literature review. *Machines*, *8*(2), 31.

Sankararaman, S. (2015). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing*, *52-53*, 228-247.

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and health management*, *1*(1), 4–23.

Tod, G., Mazaev, G., Eryilmaz, K., Ompusunggu, A. P., Hostens, E., & Van Hoecke, S. (2019). A convolutional neural network aided physical model improvement for ac solenoid valves diagnosis. In *2019 prognostics and system health management conference (phm-paris)* (pp. 223–227).

Characterizing Damage in Wind Turbine Mooring Using a Data-Driven Predictor Model within a Particle Filtering Estimation Framework

Rohit Kumar[1], Ananay Thakur[2], Shereena OA[3], Arvind Keprate[4] and Subhamoy Sen[5]

[1,2,3,5] *i4S Laboratory, Indian Institute of Technology Mandi, Mandi,175075, HP, India*
rohit373k@gmail.com
thakur07ananay@gmail.com
oa.shereena21@gmail.com
subhamoy@iitmandi.ac.in

[4] *Department of Mechanical, Electrical, and Chemical Engineering, Oslo Metropolitan University, 0166 Oslo, Norway*
arvindke@oslomet.no

ABSTRACT

Floating Offshore Wind Turbines (FOWT) represent a promising solution to renewable energy challenges, yet effective maintenance remains critical for cost management. Traditional machine learning (ML) approaches for detecting FOWT damage often rely on extensive real-world data, which can be impractical and economically unfeasible. Alternatively, stochastic filtering-based time-domain approaches leverage physical understanding through dynamic models, typically finite element models. However, these methods are hindered by excessive simulation calls within the recursive filtering frameworks. This study proposes a novel filtering-based approach that replaces the computationally intensive process model with a Deep Neural Network (DNN) surrogate, addressing the aforementioned limitations. The proposed approach utilizes synthetic data generated from the high-fidelity calibrated OpenFAST model of FOWT dynamics to train a DNN toward learning the dynamic evolution of the FOWT conditioned on the current health state. By offering a computationally efficient representation of system dynamics conditioned on health state, this approach allows for real-time damage detection and interpretable information on damage severity within a stochastic inverse estimation framework, specifically employing Particle Filtering in this study. This approach contrasts with traditional black-box ML-based methods, which typically struggle to provide interpretable information on damage characteristics. Extensive numerical investigations on damaged FOWT mooring lines demonstrate this approach's practical applicability and superiority over traditional ML-based methods. Eventually, integrating explainable ML models within the filtering framework induces promptness in detection without sacrificing transparency.

1. INTRODUCTION

Research on Floating Offshore Wind Turbines (FOWTs) has advanced significantly, reflecting their growing adoption across industries. A key challenge is maintaining safety while minimizing maintenance costs, a critical issue that persists. Structural health monitoring (SHM), particularly data-driven methods, is valued for its noise robustness and cost-effectiveness, as demonstrated by (Azimi, Eslamlou, & Pekcan, 2020).

The complex inverse problems in SHM mandate linking measurements to causes or damages. Machine Learning (ML) approaches have showcased excellent reliability and predictability across applications, yet dependence on data alone raises concerns, especially in mooring line damage detection (Avci, Abdeljaber, & Kiranyaz, 2022). In ML-based system identification for FOWTs, strategies involve extracting damage-sensitive features (DSFs) using supervised or unsupervised learning. While unsupervised methods, such as novelty detection, are effective in damage detection, they face challenges in localization and quantification (Wang, Tian, Peng, & Luo, 2018). Supervised techniques, on the other hand, require large datasets and can function as classifiers or regressors, employing algorithms such as random forest, support vector machine (SVM), and multi-layer perceptron (MLP) (Regan, Beale, & Inalpolat, 2017).

In FOWTs, the selection of appropriate DSFs for mooring line damage detection holds significant importance. Super-

vised algorithms depend on the precise representation of damage scenarios through these selected DSFs. However, it's important to acknowledge that the resulting model is essentially a black box, lacking interpretability to extract additional information not included during training. Consequently, ML-based approaches relying solely on data may not comprehensively address mooring line damage detection, underscoring the necessity for a nuanced monitoring strategy (Malekloo, Ozer, AlHamaydeh, & Girolami, 2022).

1.1. Condition Monitoring of Mooring Lines

Mooring lines are vital components in ensuring the integrity of FOWTs, influencing the optimization of support structures (Altuzarra et al., 2022). Given their significance, monitoring the health of mooring lines is essential due to potential stability implications (Aqdam, Ettefagh, & Hassannejad, 2018). While deep learning (DL) algorithms show promise in detecting damages to wind systems (Choe, Kim, & Kim, 2021), it remains imperative to understand the behavior of coupled systems under extreme conditions (Li, Le, Ding, Zhang, & Zhang, 2019), necessitating the incorporation of physics-based or physics-guided support models. Despite the prevalence of model-based and fuzzy logic approaches for mooring damage diagnosis (Jamalkia, Ettefagh, & Mojtahedi, 2016) in current literature, research on ML and DL in this domain is limited. Vibration measurements facilitate efficient identification of structural damage, aiding in damage diagnosis across various domains, including FOWTs (Farrar, Doebling, & Nix, 2001). Recent studies (Gorostidi, Pardo, & Nava, 2023) highlight the advantages of ML over model-based methods in managing large data and promptness in detection.

Traditional ML models pose significant limitations for real-life SHM due to their lack of interpretability. These black-box models, while effective at processing large amounts of data, often fail to provide meaningful insights into the underlying dynamics of the monitored system. Alternatively, stochastic filtering-based approaches, although capable of incorporating physical understanding through complex models, suffer from the computational burden of these models, making them impractical for real-time applications. However, a compromise between interpretability and efficiency can be achieved by leveraging ML techniques to create a surrogate of the conceptual model. This approach, as proposed in this study, involves replacing the computationally intensive process model with a Deep Neural Networks (DNN) surrogate. By training the DNN with real (/synthetic) data sampled (/generated) from reality (/a high-fidelity dynamic model of the system), the proposed method offers both computational efficiency and interpretability. This surrogate model can then be seamlessly integrated into a stochastic filtering framework, providing real-time damage detection with required promptness while maintaining transparency and accuracy. Thus, the study bridges the gap between conventional ML-based approaches and complex stochastic filtering methods, offering a promising solution for effective SHM in practical applications.

2. METHODOLOGY

A process model plays a major role in filtering-based methods and is often built with a Finite Element (FE) modeling approach. These models, derived from physical systems, simulate and predict the system behavior under diverse operational conditions. Despite their widespread use in evaluating civil structure conditions and detecting damage, FE models face certain challenges, such as numerical convergence issues, memory demands, and complexities in parallelization. Surrogate models such as DNNs are known for their ability to identify complex patterns swiftly and accurately, particularly in one-step-ahead time series forecasting within a data-based time-series modeling framework, and hence are one of the best choices for replacing FE models.

DNN models are faster, more adaptable, and require comparatively less simulation effort than traditional FE models. DNNs also offer other computational advantages like scalability, capturing nonlinear data relationships efficiently, thereby enabling quick and accurate predictions without the need for iterative solutions. This streamlined approach accelerates development and reduces costs associated with model building and simulation.

Typical health assessment problems, addressed with conventional DNN models, require extensive datasets correlating response inputs with damage locations and severity labels to achieve comprehensive accuracy. To mitigate this issue, instead of correlating the response to its corresponding damage labels, the underlying dynamics are learned within a DNN framework utilizing a response time series of consecutive time steps as input-output pairs. The input is additionally augmented with the health states of the system to render the prediction conditional on the health state. Once trained, the unobserved health state can therefore be observed with the DNN network in terms of response. Filter-based estimation methods further leverage this mapping to inversely estimate the health state inferred from the measured response time history. However, before that, exploring the sensitivity of the prepared DNN models compared to traditional FEM-based models is essential to justifying their adoption over costlier FEM-based predictors.

The subsequent discussion focuses on a DNN model trained using simulations from an FEM-based model, synthesized from software like OpenFAST designed for FOWTs. Actual sampled responses will replace simulated ones for real-world implementation.

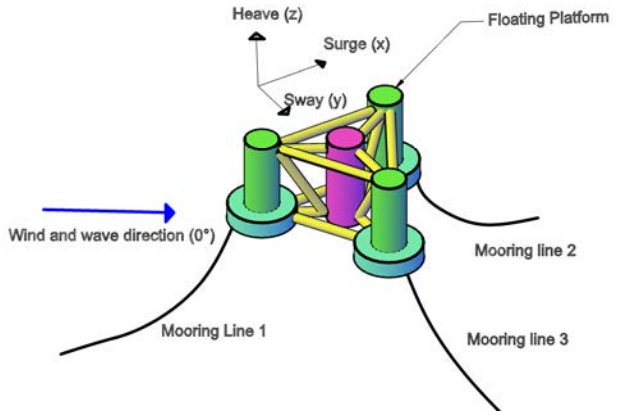


Figure 1. OC4 Semisubmersible floater along with catenary mooring system for NREL's 5MW wind turbine.

2.1. OpenFAST Model

Ensuring the DNN model undergoes thorough training necessitates a substantial volume of data. However, due to the limited prevalence and operational scope of FOWTs, obtaining a satisfactory quantity of authentic data is challenging. Consequently, we address this requirement by generating the requisite data through high-fidelity software capable of multi-physics simulation. In this investigation, we employ an NREL 5 MW Wind turbine model affixed to an OC4 semi-submersible model, as depicted in Figure 1. This configuration incorporates a catenary mooring system with three mooring lines fastened at 120° angles (Robertson et al., 2014); detailed mooring line specifications are provided in Table 1. The specifications for the 5 MW reference turbine are outlined in (Jonkman, Butterfield, Musial, & Scott, 2009). Additionally, data across varying sea states are simulated, characterized by wide-band operational scenarios involving a significant wave height (H_s) of 6 m and a Peak period (T_p) of 10 sec, assuming the turbine operates under a constant, steady wind speed of 8 m/s while in full operational mode.

Table 1. Mooring lines details

Diameter (m)	Mass density (kg m^{-1})	Axial stiffness (N)	Unstretched length (m)
0.0766	113.35	7.5903×10^8	835.35

In OpenFAST, various modules perform distinct functions. For instance, the HydroDyn module adopts a hybrid approach to handle hydrodynamic loads on the platform, merging diffraction theory with the Morison equation. The AeroDyn module uses blade element momentum (BEM) theory to manage aerodynamic loads. MoorDyn oversees loads related to mooring lines through the lumped mass method. ElastoDyn addresses structural and gravitational loads, while the InflowWind module supplies essential wind output. These modules are interconnected, collaborating to simulate the desired responses.

Table 2. Mooring Line's Damage (D) and Healthy (H) scenarios

Cases	Mooring Line's Damage Scenarios		
	Line 1 (k_1)	Line 2 (k_2)	Line 3 (k_3)
Case 1	H	H	H
Case 2	D	H	H
Case 3	H	D	H
Case 4	H	H	D
Case 5	D	D	H
Case 6	H	D	D
Case 7	D	H	D
Case 8	D	D	D

The DNN network underwent training to support the particle filter, capturing six distinct responses: the displacement and velocities of the floating platform in three directional axes. Response data was simulated over a duration of 3600 seconds, sampled at a frequency of 40 Hz. Variations in response resulting from alterations in the material properties of the mooring line were documented, leading to the simulation of different scenarios corresponding to various combinations of mooring line damage. Specifically, damages to the mooring lines were introduced as reductions in axial stiffness. For each scenario, 240 samples were simulated, resulting in a total of 1920 samples across 8 distinct cases as outlined in Table 2. Each case encompasses three discrete levels of damage for each mooring line, representing reductions of 10%, 15%, and 20% in axial stiffness (k_1, k_2, and k_3) of the mooring material.

2.2. Deep Neural Network (DNN)

The DNN model was subsequently developed to learn from data generated by OpenFAST. To predict the displacements and the velocities of a floating platform at the next ($k + 1^{th}$) sampling time step using the current time step (k^{th}) responses along with the health states as the input, the DNN utilizes a technique called back-propagation, wherein the prediction errors are propagated backward through the network, allowing the model to adjust its internal parameters, to improve future predictions. The architecture for the designed DNN model is provided in Table 3.

Table 3. Characteristics of the DNN model

DNN Architecture		
Layers	**Activation Function**	**Nodes**
Input Layer	-	9
Hidden Layer - 1	ReLU	128
Hidden Layer - 2	ReLU	64
Output Layer	Linear	6
Hyperparameters		
Optimizer		Adam
Epochs		1000
Learning Rate		10^{-6}

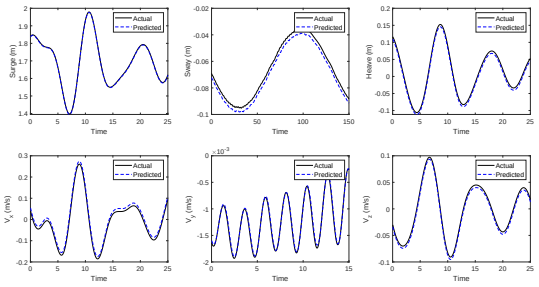


Figure 2. Actual and one-time step ahead predicted response comparison

The DNN architecture comprises an input, an output, and two hidden layers. The model is trained using 70% of the dataset allocated as training data. The remaining, 20% of data is used as a validation set to optimize hyper-parameters (activation function, learning rate, batch size) using RMSE and MAE metrics. Subsequently, testing is conducted using the remaining 10% of data. To ensure the required architecture for datasets, ReLU (Rectified Linear Unit) activation and linear activation functions are used in the hidden layers and output layer, respectively. The Adam optimizer, with a learning rate of 10^{-6} for 1000 epochs, was used for model optimization.

The model is tested on 10% of datasets. The model performed well in the provided regression task with good Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values, as shown in Table 4. Further, a comparison of one-time step-ahead predictions and actual simulated responses is shown in Figure 2.

Table 4. Trained DNN model's accuracy indices for testing data

	RMSE	**MAE**	**MBE**
Surge (m)	0.0014151	7.7249×10^{-5}	-9.0×10^{-5}
Sway (m)	0.0001885	9.1052×10^{-6}	9.0×10^{-5}
Heave (m)	0.0007831	2.5640×10^{-5}	-1.77×10^{-4}
V_x **(m/s)**	0.0022982	0.0015	-3.245×10^{-3}
V_y **(m/s)**	3.4311×10^{-5}	2.8587×10^{-5}	-1.8×10^{-5}
V_z **(m/s**	0.0006027	0.0003	9.8×10^{-5}

The performance of the DNN model, in predicting various responses, has been evaluated based on the provided accuracy indices. The RMSE values ranging from 3.4311×10^{-5} to 0.0022982 indicate relatively low prediction errors across different parameters, suggesting the model's capability to make accurate one-time step-ahead predictions. Likewise, the MAE values, ranging from 9.1052×10^{-6} to 0.0015, demonstrate the model's ability to predict parameter values with small deviations from the true values. Despite some biases observed in the Mean Bias Error (MBE) values, their magnitudes are relatively small, indicating an unbiased prediction by the ANN model. Overall, these accuracy indices suggest that the ANN model performs well in predicting the parameters of interest, making it useful for particle filters.

2.3. DNN-Particle filter

Particle filter, which typically approaches the Sequential Monte Carlo (SMC) method, is a powerful technique used for state estimation in nonlinear and non-Gaussian systems. The computational difficulties associated with scenarios where the state-space model is complex and the direct analytical solutions are intractable are handled effectively by PF-based algorithms.

The central idea behind a particle filter is that the posterior distribution ($p(x_{k-1}|R_{k-1})$) of the system state could be represented using a set of weighted samples, called particles $\{x_k^{(i)} : i = 1, 2, 3, \ldots, N\}$. These particles evolve according to the system dynamics and are updated based on their likelihood against measurements arriving sequentially in time. The filter approximates the posterior distribution by propagating particles through the system dynamics and adjusting their weights based on the likelihood of observed measurements.

The key steps in a particle filter algorithm include prediction, measurement update, and resampling. In the prediction step, particles are propagated forward in time according to the system dynamics, incorporating process noise, if present. In the measurement update step, particles are weighted based on their consistency with the observed measurements, calculated using the likelihood function. Upon receiving measurements, the probability of each sample from the previous time step is evaluated, and the normalized weight of each sample is determined using Eq. (1).

$$a_i = \frac{p(\Phi_k|\widetilde{x}_k^{(i)})}{\sum_{j=1}^{N} p(\Phi_k|\widetilde{x}_k^{(j)})} \tag{1}$$

Each sample $\widetilde{x}_k^{(i)} : i = 1, 2, 3, \ldots, N$ forms a discrete distribution with probability mass a_i associated with element i. Resampling is then performed to prevent particle degeneracy by replicating particles with higher weights and eliminating those with lower weights, redistributing the particle set to high-likelihood areas. The above process helps ensure a diverse representation of the posterior distribution. Resampling the discrete distribution N times creates new samples, weighted based on their likelihood against the observed data. Particle filters can handle nonlinear, non-Gaussian systems without linearization but may suffer in high-dimensional spaces due to the curse of dimensionality, requiring careful parameter tuning like the number of particles to balance efficiency and accuracy.

In essence, particle filters offer a flexible and effective framework for state estimation in nonlinear, non-Gaussian systems, serving as valuable tools in robotics, target tracking, and fi-

nancial modeling.

2.3.1. Parameter Estimation via PF

Within the PF framework, the adopted damage attributes, posed as parameters θ_k, are defined with a set of N_p independent parameter particles $\boldsymbol{\xi} = [\xi_{k-1}^1, \xi_{k-1}^2, \cdots, \xi_{k-1}^{N_p}]$ (?, ?). These particles, each representing a possible state of the system, are used to propagate system uncertainty over time through a process model. Subsequently, the propagated particles are evaluated against available measurements using the measurement model to compute their likelihood. This likelihood, when combined with the prior likelihood of particles, forms the posterior distribution. Finally, new particle samples are drawn from this posterior distribution to be utilized in the next iteration of the process.

The process model for this particle filter demonstrates the evolution of the parameter vector $\theta_k : \mathbb{R}^p$ to be estimated using a random walk model as follows:

$$\theta_{k+1} = \theta_k + u_k \tag{2}$$

θ_k signifies the health states, typically material properties, stiffness, or health indices, through which damage can be characterized. This model allows the parameter states to evolve in time to converge to their respective true values. Within the particle filtering framework, this uncertainty propagation is achieved by the time updating of several particles through the process model, along with the associated uncertainty u_k that has been modeled as a stationary white Gaussian noise (SWGN) of covariance Q_k.

Further, with each time step k, the evolution of the particle, ξ_{k-1}^j, is essentially represented by a random perturbation around its current position. A Gaussian blur ($\mathbb{N}(\delta\xi_k, \sigma_k^\xi)$) is additionally applied to ξ_{k-1}^j with a shift $\delta\xi_k = (1-\alpha)\bar{\xi}_{k-1}$ and a spread of σ_k^ξ [1]. The turbulence in the particle estimation is effectively managed through the implementation of hyperparameter α, which attempts to re-center the particles towards their mean ($\bar{\xi}_{k-1}$) as,

$$\xi_k^j = \alpha\xi_{k-1}^j + \mathcal{N}(\delta\xi_k, \sigma_k^\xi) \tag{3}$$

After propagation, each parameter particle undergoes observation against available measurements utilizing the DNN model. This model can map current responses and parameters to responses at the subsequent time step. The further mapping of the response at the next step to its corresponding available response is not explicitly elaborated here and is collectively incorporated within the measurement function $h(\bullet)$. The measurement model is defined as follows:

$$y_{k+1} = h_k(x_k, \theta_{k+1}, v_k) \tag{4}$$

In this context, h_k therefore utilizes the DNN surrogate of the FEM model trained with simulated synthetic response data. The DNN surrogate predicts the responses at the next time steps, some of which are observed as measurements y_{k+1}. v_k denotes measurement noise, modeled as another SWGN with covariance R_k. In the current FOWT monitoring scenario, the parameter vector encompasses the stiffness parameters of three mooring catenaries (k_1, k_2, and k_3).

Using this process and measurement model, the particle filter estimates the posterior of parameter particles, and the particle mean leads to the estimate of the particle filters. Due to space constraints, a detailed description of the particle filter is not provided here. Interested readers are encouraged to refer to the extensive literature available in this field.

2.3.2. Particle update and particle approximation

Next, the likelihood, $\mathcal{L}(\xi_k^j)$ for each j^{th} particle is computed using the corresponding innovation mean and covariance. These likelihoods are further convoluted with the prior weights $w(\xi_{k-1}^j)$ to estimate the corresponding posterior $w(\xi_k^j)$.

$$w(\xi_k^j) = \frac{w(\xi_{k-1}^j)\mathcal{L}(\xi_k^j)}{\sum_{j=1}^{N_p} w(\xi_{k-1}^j)\mathcal{L}(\xi_k^j)} \quad \text{with}$$
$$\mathcal{L}(\xi_k^j) = \left((2\pi)^n \sqrt{|\mathbf{R}_k|} \right)^{-1} e^{-0.5 i_k^{j\,T} \mathbf{S}_k^{-1} i_k^j} \tag{5}$$

With these updated weights, the particle approximations for the parameters are then estimated as:

$$\xi_{k|k} = \sum_{j=1}^{N_p} w(\xi_k^j)\xi_k^j \tag{6}$$

3. Results and Discussion

The integration of a DNN-based process model within a particle filtering environment underwent testing for two numerical case studies under two different operational conditions (refer Case 2 and Case 8). Under operating condition case 2, the numerical experiment additionally considers 10% damage in the first mooring line alone ($k_1 = 0.9k$), and no damage in other mooring lines, while under operating condition case 8, the numerical experiment assumes 20% damage in each of all the three mooring lines. To simulate real-world conditions, the observation vector was contaminated with 1% and 5% Gaussian noise. The objective was to assess the efficiency and robustness of the proposed detection approach in estimat-

[1] $A + B\mathbb{N}(\mu, \sigma)$ means $A + Bz$, where z follows $\mathbb{N}(\mu, \sigma)$

Algorithm 1 DNN-Particle Filter

Inputs:
N: Number of particles
$x(0)$: Initial state estimate
$p(x(0))$: Initial state probability distribution
$f(x_k, u_k)$: System dynamics function
$h_k(x_k, \theta_k)$: Measurement model function
$DNN(x_k, \theta_k)$: DNN surrogate model for measurement update STATE σ_ξ: Standard deviation for particle perturbation
α: Hyper-parameter for particle re-centering
Q_k: Process noise covariance
R_k: Measurement noise covariance
Outputs:
Estimated state posterior: $p(x_k|z_1 : k)$ (represented by particles)
Estimated parameter vector: θ_k
Initialization:
for $i = 1$ **to** N **do**
    Sample initial state: $x_0^{(i)} = \text{Sample\_From}(p(x(0)))$
    Initialize parameter particles:
    $\theta_0^{(i)} = \text{Random\_Vector}()$
    Initialize weights: $w_0^{(i)} = 1/N$
end for
Main Loop:
for $k = 1$ **to** T (number of time steps) **do**
    Prediction Step:
    for $i = 1$ **to** N **do**
        Propagate particle state: $x_k^{(i)} = f(x_{(k-1)}^{(i)}, u_k)$
        Perturb parameter particle:
        $\theta_k^{(i)} = \alpha\theta_{(k-1)}^{(i)} + \text{N}(0, \sigma_\xi)$
    end for
    Measurement Update Step:
    for $i = 1$ **to** N **do**
        Calculate innovation:
        $\text{innovation} = y_k - h_k(x_k^{(i)}, \theta_k^{(i)})$
        Calculate likelihood:
        $\mathcal{L}(\xi_k^j) = \left((2\pi)^n\sqrt{|\mathbf{R}_k|}\right)^{-1} \exp\left(-0.5(\boldsymbol{i}_k^j)^T\mathbf{S}_k^{-1}\boldsymbol{i}_k^j\right)$
        Update weight based on likelihood:
        $w_k^{(i)} = w_{(k-1)}^{(i)} \cdot \mathcal{L}(\xi_j^k)$
    end for
    Normalize weights: $w_i = \frac{w_i}{\sum_{j=1}^N w_j}$
    Resampling Step:
    Perform resampling to generate new particles and parameters based on weights
    Update particles and parameters based on resampling results
end for

ing the stiffness parameters (k_1, k_2, and k_3) of the mooring lines across varying noise levels.

Results indicate that, regardless of the noise level, the integrated model effectively estimated the states and stiffness parameters with sufficient accuracy. Figures 3, 4, 5, and 6 demonstrate the proposed framework's parameter estimation capabilities even in the presence of noise. However, an increase in noise led to decreased accuracy in parameter estimation. Higher noise levels introduced greater ambiguity into the observation vector, resulting in increased inaccuracy in the estimation process. The accuracy of the estimated parameters is quantified in terms of RMSE, with lower RMSE values indicating better accuracy.

Although all 8 scenarios in Table 2 were studied with the proposed estimation framework, only the results corresponding to Case 2 and Case 8 are presented in the paper for the numerical demonstration due to technical limitations.

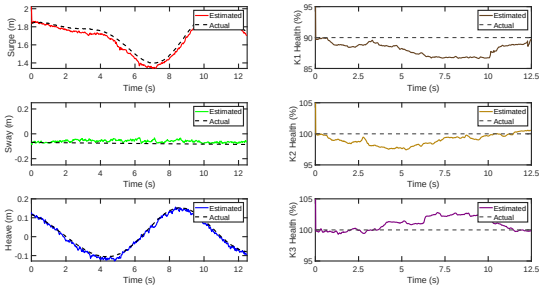


Figure 3. Actual and estimated states (left) and stiffness (right) parameters for Case 2 with 1% noise.

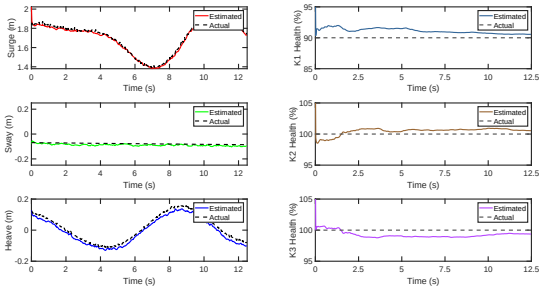


Figure 4. Actual and estimated states (left) and stiffness (right) parameters for Case 2 with 5% noise.

4. Conclusion

The initial findings suggest that the DNN-particle filter holds promise as a means to enhance the reliability and efficiency of mooring line monitoring by integrating an ML-based predictor model instead of a high-fidelity FEM model. By leveraging synthetic data generated from a calibrated OpenFAST model of FOWT dynamics, the DNN-particle filter offers a

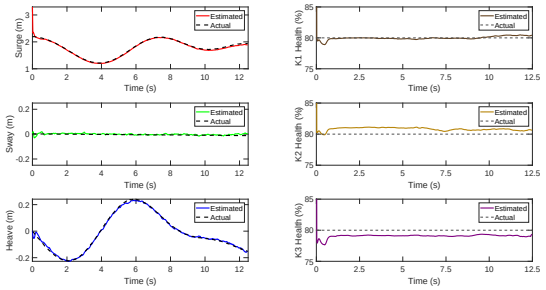


Figure 5. Actual and estimated states (left) and stiffness (right) parameters for Case 8 with 1% noise.

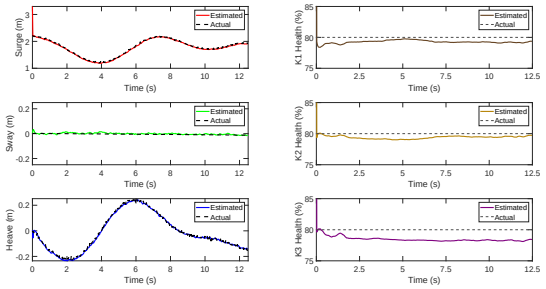


Figure 6. Actual and estimated states (left) and stiffness (right) parameters for Case 8 with 5% noise.

computationally efficient representation of system dynamics, enabling real-time damage detection and interpretable information on damage severity within a stochastic inverse estimation framework. This contrasts with traditional black-box ML-based methods, which often struggle to provide interpretable information on damage characteristics.

The DNN-particle filter's data processing capabilities maximize resource allocation by focusing efforts on critical areas identified by the model. By doing so, it enhances the promptness of the detection algorithm without sacrificing accuracy and transparency. This development has the potential to usher offshore operations into a new era of durability and resilience by ensuring the integrity of mooring lines and streamlining operating procedures. Ultimately, it has the potential to enhance the safety and longevity of offshore operations by effectively managing damage and noise.

REFERENCES

Altuzarra, J., Herrera, A., Matías, O., Urbano, J., Romero, C., Wang, S., & Guedes Soares, C. (2022). Mooring system transport and installation logistics for a floating offshore wind farm in lannion, france. *Journal of marine science and engineering*, *10*(10), 1354.

Aqdam, H. R., Ettefagh, M. M., & Hassannejad, R. (2018). Health monitoring of mooring lines in floating structures using artificial neural networks. *Ocean Engineering*, *164*, 284–297.

Avci, O., Abdeljaber, O., & Kiranyaz, S. (2022). An overview of deep learning methods used in vibration-based damage detection in civil engineering. In *Dynamics of civil structures, volume 2: Proceedings of the 39th imac, a conference and exposition on structural dynamics 2021* (pp. 93–98).

Azimi, M., Eslamlou, A. D., & Pekcan, G. (2020). Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review. *Sensors*, *20*(10), 2778.

Choe, D.-E., Kim, H.-C., & Kim, M.-H. (2021). Sequence-based modeling of deep learning with lstm and gru networks for structural damage detection of floating offshore wind turbine blades. *Renewable Energy*, *174*, 218–235.

Farrar, C. R., Doebling, S. W., & Nix, D. A. (2001). Vibration–based structural damage identification. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *359*(1778), 131–149.

Gorostidi, N., Pardo, D., & Nava, V. (2023). Diagnosis of the health status of mooring systems for floating offshore wind turbines using autoencoders. *Ocean Engineering*, *287*, 115862.

Jamalkia, A., Ettefagh, M. M., & Mojtahedi, A. (2016). Damage detection of tlp and spar floating wind turbine using dynamic response of the structure. *Ocean Engineering*, *125*, 191–202.

Jonkman, J., Butterfield, S., Musial, W., & Scott, G. (2009, 2). Definition of a 5-mw reference wind turbine for offshore system development. Retrieved from `https://www.osti.gov/biblio/947422` doi: 10.2172/947422

Li, Y., Le, C., Ding, H., Zhang, P., & Zhang, J. (2019). Dynamic response for a submerged floating offshore wind turbine with different mooring configurations. *Journal of Marine Science and Engineering*, *7*(4), 115.

Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. *Structural Health Monitoring*, *21*(4), 1906–1955.

Regan, T., Beale, C., & Inalpolat, M. (2017). Wind turbine blade damage detection using supervised machine learning algorithms. *Journal of Vibration and Acoustics*, *139*(6), 061010.

Robertson, A., Jonkman, J., Masciola, M., Song, H., Goupee, A., Coulling, A., & Luan, C. (2014). *Definition of the semisubmersible floating system for phase ii of oc4* (Tech. Rep.). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Wang, P., Tian, X., Peng, T., & Luo, Y. (2018). A review

of the state-of-the-art developments in the field monitoring of offshore structures. *Ocean Engineering*, *147*, 148–164.

Comparison among Machine Learning Models Applied in Lithium-ion Battery Internal Short Circuit Detection

ZiHong Zhang[1], Mikel Arrinda[2], and Jon Perez[3]*

[1,2,3] *CIDETEC, Basque Research and Technology Alliance (BRTA), Po. Miramón 196, 20014 Donostia-San Sebastián, Spain*

zzhang@cidetec.es
marrinda@cidetec.es
jonperez@cidetec.es

ABSTRACT

The world is experimenting a decarbonization process, mainly through lithium-ion-based solutions. Nonetheless, catastrophic events have negatively affected the social acceptance of lithium-ion-based solutions. One of the most interesting projects regarding catastrophic event prevention is the internal short-circuit detection. This paper proposes to detect it using different machine-learning algorithms such as random forest and combination of random forest with neural network-based algorithms through time-instant classification and historical feature classification. The hyper-parameters have been optimized through grid-search. The selected algorithms have been trained thanks to synthetically generated data using a first-order electrical equivalent circuit model. The performance of the generated models has been verified and compared thanks to testing and validation data sets taken from the synthetically generated data. Afterward, the most accurate internal short circuit detection algorithm was selected and validated through laboratory-level data. The selected cell in this study is SLPB526495HE, a pouch cell of 3.7Ah. The generated data are time series of voltage and current, which are the variables that will be available in a real application. The results demonstrate an accuracy above 90% in detecting an internal short circuit in the most interesting cases. The validation with laboratory data has shown that an accuracy of 90% can be achieved. This paper provides learned lessons on the process of developing the internal short circuit detection machine-learning model, highlighting the potential they possess to detect accurately internal short circuits.

1. INTRODUCTION

Lithium-ion batteries have been acclaimed for their high energy density, low self-discharge rates, and environmental compatibility since a decade ago (Diouf & Pode, 2015). This battery technology has emerged as a key component in global decarbonization strategies, finding extensive application in diverse energy storage systems such as electric vehicles and smart grids (Zubi et al., 2018). Despite their notable advantages, safety concerns, particularly those stemming from internal short circuits (ISC) leading to thermal runaway, remain a primary impediment to their broader adoption(Zhan et al., 2023). Such incidents can result in battery fires or even explosions, leading to grave consequences. Thermal runaway is often initiated by ISC events (Ren et al., 2021), and the detection of such events poses significant challenges, especially during their incipient stages.

ISC faults often begin with mild severity, starting with high resistance values, which decrease as the fault progresses. In this process, the voltage, current, and State of Charge (SoC) of normal batteries and those with varying ISC resistance values exhibit similar characteristics during charging and discharging processes. This similarity significantly complicates the diagnosis of early-stage ISC faults, leading to researchers to apply data-driven algorithms (Zhang et al., 2021). The application of data-driven or machine learning (ML) algorithms in ISC fault detection can be categorized further into two main types: unsupervised and supervised learning.

Unsupervised learning methods train fault detection models using data generated during the charging and discharging processes of normal batteries. These methods identify potential anomalies by defining deviation measures between normal and abnormal data, deciphering the standard patterns within the data, and employing specific decision rules. Typical algorithms include Support Vector Machine (SVM) (Chatterjee et al., 2023), Relevance Vector Machine (RVM) (Xie et al., 2020), Kernel Principal Component Analysis (KPCA) (Schmid & Endisch, 2022), and Isolated Forest (Cheng et al., 2023), which have demonstrated effective anomaly detection capabilities in various scenarios. Nonetheless, unsupervised learning algorithms have their limitations. Particularly when the abnormal data closely resemble the normal data with no significant distributional

differences, as it is our case. These methods may struggle to distinguish between them accurately. This is why, it is challenging for unsupervised learning algorithms to detect these subtle anomalies.

Supervised learning methods utilize pre-labeled datasets for training, differentiating like this battery performance data under normal operations and abnormal conditions in the training process. Through training, these models are adept at distinguishing between normal battery behavior and potential fault signals. Typical algorithms include Random Forest (RF) (naha et al., 2020), Convolutional Neural Network (CNN) (Yang et al., 2022) and Long Short-Term Memory (LSTM) (Wang et al., 2023). Nonetheless, it is not clear which should be the one to be applied. In light of this, our research aims to explore and apply various supervised learning algorithms to help researchers find the most suitable algorithms for ISC detection.

This paper proposes the development and comparison of ISC detection supervised learning algorithms, enhancing the detection capabilities at early stages of ISC. This study seeks to provide an ISC detection algorithm selection background to fellow researchers and boost the reliability and safety of lithium-ion batteries in operational contexts.

This paper is structured as follows. The data generation is detailed in section 2. The ISC detection methods are described in section 3. The hyperparameter tuning process undergone in this paper is placed in section 4. The results are shown in section 5. The discussion is done in section 6 and the conclusions are drawn in section 7.

2. DATA GENERATION

The selected battery is SLPB526495HE. The synthetically generated data has been generated with a first-order equivalent electric model. The experimental data has been generated in laboratory testing facilities. During the training and testing phases, only virtual datasets were utilized to develop the models. In the validation phase, experimental datasets were additionally incorporated. This approach was adopted to evaluate the performance of the models trained on virtual datasets in real-world scenarios.

This study focuses on charging data. In practical scenarios, battery discharging conditions are highly complex, whereas the charging scenarios are relatively monotonous. Therefore, we chose charging data to train the model for detecting ISC anomalies during the charging phase.

The operational conditions of the generated data are the same for the synthetically generated one and the one generated through laboratory tests: an ambient temperature of 25°C and a charge process at constant charge mode from 1% SoC to the maximum voltage value.

The extracted feature data during this process included the battery voltage and the battery's current voltage increment relative to its voltage before charging an amount equivalent to 1% of its nominal capacity, denominated as voltage difference, see Eq. (1).

$$V_{diff\_k} = V_k - V_{k-1\%c} \quad (1)$$

2.1. Virtual dataset

The virtual data set used for model training and testing was generated by a first-order electric equivalent circuit model (Arrinda, Oyarbide, Macicior, Muxika, et al., 2021) for the SLPB526495HE battery, as shown in Figure 1.

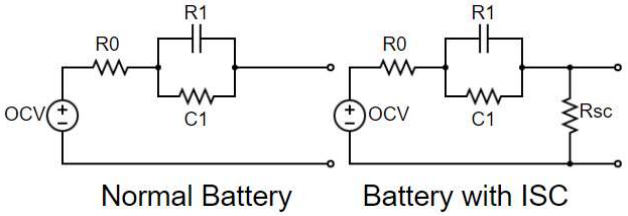


Figure 1: Equivalent electric circuit models.

The parameters of the model were obtained by conducting specific modeling tests on the SLPB526495HE battery, see Figure 2.

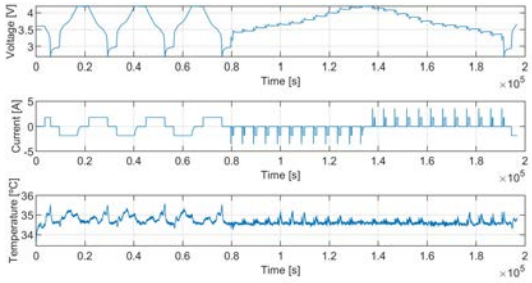


Figure 2: Modeling tests. A capacity and pulse based impedance characterization and OCV characterization test is performed at controlled ambient temperature.

The built model was run to obtain data from a normal battery and a battery with different level of ISC faults (a total of 21 stages): 5Ω, 50Ω, 100Ω, 150Ω, 200Ω, 250Ω, 300Ω, 350Ω, 400Ω, 450Ω, 500Ω, 550Ω, 600Ω, 650Ω, 700Ω, 750Ω, 800Ω, 850Ω, 900Ω, 950Ω and 1kΩ.

2.2. Experimental dataset

The experimental Data Set used for the model validation was generated in the laboratory by cycling the cell with constant current charging tests from 1% SOC to maximum voltage value, see Figure 3. The reference test has been tested only with the cell. The ISC has been emulated by connecting an external bleed resistor of 10 Ω and performing the charge.

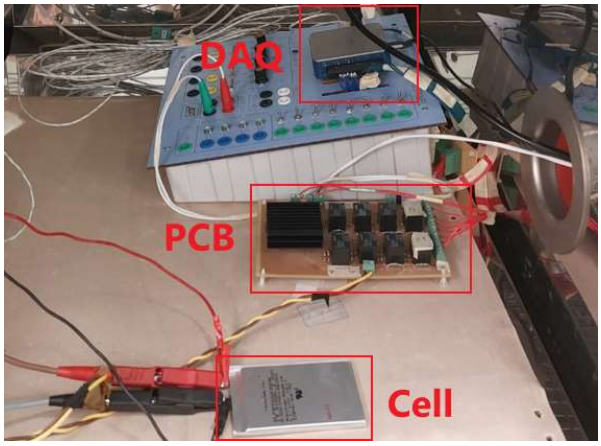


Figure 3: Experimental setup. Within a controlled temperature chamber, a battery cell is interfaced with a Printed Circuit Board (PCB), which incorporates various resistors to simulate an Internal Short Circuit (ISC) phenomenon. This setup is further connected to a Data Acquisition System (DAQ) for comprehensive data collection and monitoring.

3. DETECTION METHODS

This paper presents and compares various ISC detection solutions based on different ML models. These solutions can be divided into two main categories according to the data types they utilize: the Instantaneous Feature-based Method and the Historical Feature-based Method.

3.1. Instantaneous Feature-based Method

As illustrated in Figure 4, the ISC detection solutions under the Instantaneous Feature-based Method determine whether the battery is in a normal state or experiencing an ISC anomaly by analyzing the feature data at every single moment and treating it as a binary classification task at every moment. The ISC detection solutions proposed in this paper that are under the Instantaneous Feature-based Method category are the RF solution and the RF combined with Multilayer Perceptron (RF+MLP) solution.

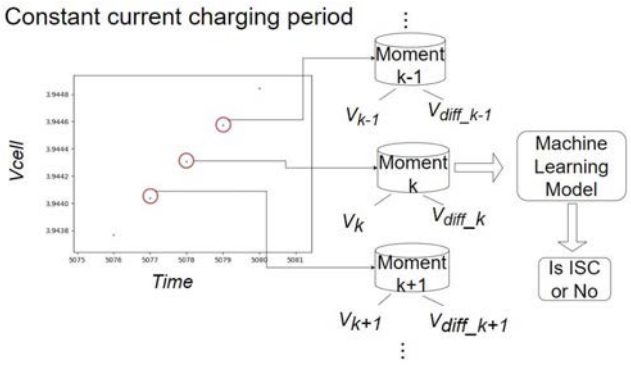


Figure 4: Instant-based methods' main concept diagram. Instant-based methods treat each moment k as an independent data point, characterized by two features: the cell voltage at time k (Vk) and the voltage difference (Vdiff_k) at time k. A machine learning model is then employed to classify each time point as either a normal or an ISC label.

3.1.1. Data treatment

A dataset for the charging process on a battery without an ISC was gathered with no ISC labels. A dataset of battery charging data that reflects the 21 stages of ISC fault conditions has been generated with ISC anomaly labels. The dataset with ISC anomaly labels has a significantly higher volume of data compared to the one with no ISC labels. To address this issue, a down-sampling method was adopted to balance the label distribution in the training dataset.

The specific down-sampling process involves randomly selecting data from the dataset with ISC anomaly labels. The total number of data points is the same in both data sets, 4000. The data with ISC anomaly labels is evenly taken from all the simulated cases. This method ensures the consistency of the total volume of ISC anomaly data with normal data and guarantees a balanced sampling quantity of different stages of ISC anomaly data generated from different short-circuit resistance values.

The final step in data handling involves splitting the dataset obtained through down-sampling into a training set and a test set based on an 80% to 20% ratio. This approach allows the model to train on a substantial portion of the data while retaining a separate subset for evaluation, ensuring that the model's performance can be accurately assessed.

3.1.2. Random Forest (RF)

The RF classifier, as a widely used ML model for classification tasks, leverages ensemble learning techniques to enhance the accuracy and stability of predictions. This model employs bootstrapping to draw multiple subsets of samples with replacements from the original training dataset and randomly selects subsets of features during the construction of each decision tree. In classification tasks, RF makes the final decision by aggregating the predictions from all its decision trees, adopting the class supported by the majority of the trees as the prediction outcome.

3.1.3. Randon Forest with Multilayer Perceptron (RF+MLP)

The RF+MLP combines the RF and MLP to detect the presence of ISC phenomena in batteries. The main workflow consists of training first a RF classifier to be used to predict the data. Then, the prediction results from each decision tree within the RF classifier are used as new input features of the MLP. This approach aims to leverage the MLP to learn the relationships between decision trees, thereby enhancing the

model's ability to distinguish between data with ISC and data without ISC.

3.2. Historical Feature-based Method

The historical feature-based method for detecting ISC events utilizes a sliding window technique. Starting from the (n+1)th time point, it combines the feature data of that moment with the feature data from the preceding n moments to construct a time series window. The window then slides forward, step by step, continuing this process to generate a series of time series window data, see Figure 5. Subsequently, deep learning models specifically designed for time series classification are applied to distinguish between ISC event data and no ISC data.

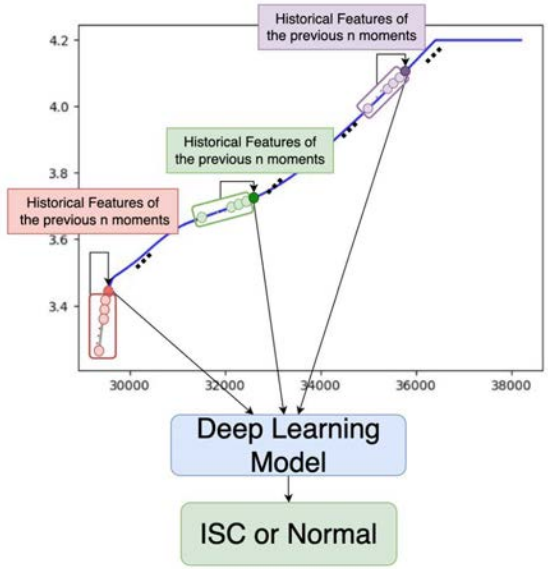


Figure 5: Historical feature-based methods' main concept diagram.

3.2.1. Data treatment

A series of data preprocessing steps are necessary. Initially, the sliding window technique is applied to transform the charging data of batteries without ISC and the charging data of batteries with ISC anomaly using different short-circuit resistance values. Starting from the (n+1)th moment, the feature data of each moment and its preceding n moments are combined to form individual 3x(n+1) dimensional time series windows. Here, 3 represents the number of features:

- The battery voltage at each moment.
- The voltage difference or the voltage increment of the battery at each moment relative to its voltage before charging an amount equivalent to 1% of its nominal capacity.
- The probability that the current moment might correspond to ISC data as determined by the Random Forest classifier analyzing the current battery voltage at each moment and the voltage difference.

Subsequently, down-sampling of ISC time series window data is performed as in the data treatment performed for the instant-based methods to balance the data label distribution and prevent data bias issues during the training process.

Unlike the RF algorithm, neural network models typically require data normalization prior to training. This normalization accelerates model convergence, prevents issues with vanishing or exploding gradients, and enhances the model's generalization capability to new data. Common data normalization methods include the min-max normalization and the Z-score normalization (Patro & sahu, 2015).

The min-max normalization method adjusts the scale of the data so that all features have values ranging between 0 and 1. Specifically, for each feature, this is achieved by subtracting the minimum value of that feature from each value, then dividing by the difference between the maximum and minimum values of that feature, Eq. (2).

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

The Z-score normalization, also known as standard score normalization, normalizes the data by subtracting the mean of each feature from its values and then dividing by its standard deviation, resulting in a dataset with a mean of 0 and a standard deviation of 1, Eq. (3).

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{3}$$

Beyond the normalization methods, the choice of normalization strategy is crucial and can be based on one of the considered normalization processes: normalization-by-moment, normalization-by-feature, and normalization-by-window.

The normalization-by-moment strategy involves normalizing the values of all features at each specific moment. It treats each point in time independently, adjusting the features across all samples at that particular moment to conform to the chosen normalization scale. This approach is useful when the relative magnitudes of features at each moment are important for the model to recognize patterns over time.

The normalization-by-feature strategy operates on each feature across all moments. It normalizes the values of a single feature over the entire dataset, ensuring that the feature's values are on the same scale across all time points. This is particularly beneficial when you want the model to understand the behavior of each feature independently across time, emphasizing the feature's overall distribution without the influence of varying scales.

The normalization-by-window approach treats all the data within a sliding window as a whole for normalization purposes. Each window is normalized independently, meaning that the scale of the features is adjusted within the context of that window. This strategy is useful when the relationship between features within each window is critical to identifying patterns, and it aims to preserve the internal dynamics of each time window.

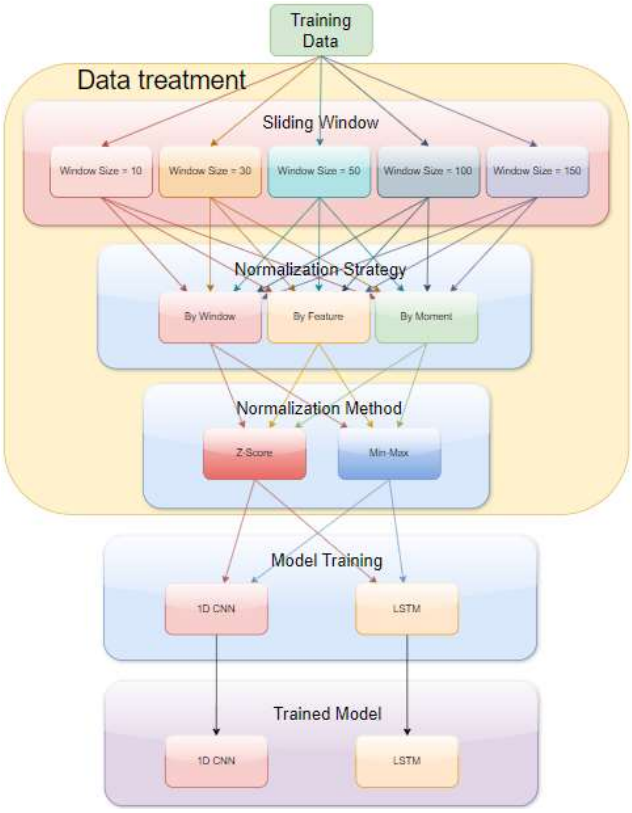


Figure 6: Data treatment processing selection diagram.

As a result, six distinct data preprocessing schemes have been developed based on the aforementioned. To identify the most suitable data processing approach for various models and to find the optimal sliding window size, an experimental workflow was designed as follows (Figure 6):

- Experiment with window sizes equals to 10, 30, 50, 100, 150 are evaluated.
- The six different data preprocessing schemes are applied to each window size experiment, resulting in 30 different data processing configurations.
- Train 1D CNN and LSTM networks using the processed data.
- Evaluate the performance of these models through cross-validation to determine the most suitable data processing method and window size for each model type.

3.2.2. Random Forest with Convolutional Neural Network (RF+CNN)

1D CNN algorithms are frequently employed for processing sequential data, such as time series data. A 1D CNN processes input data through a series of specific layers to extract useful features for classification or other tasks. The fundamental architecture of a simple 1D CNN consists of an input layer, convolutional layer, activation function, pooling layer, fully connected layer, and output layer. In the context of time series classification tasks, the input layer initially receives the raw data. This is followed by the convolutional layer, where multiple kernels slide across all features of the data to perform convolution operations and extract features, which are subsequently subjected to an activation function. The pooling layer then reduces the dimensionality of the feature maps, decreasing the volume of data that needs to be processed. Finally, the fully connected layer and the output layer classify the previously extracted features, producing the final outcome.

3.2.3. Random Forest with Long short-term memory (RF+LSTM)

LSTM networks are a specialized type of Recurrent Neural Networks (RNNs) particularly suited for classifying, processing, and predicting based on time series data. LSTMs are adept at addressing issues of vanishing or exploding gradients, which are common with traditional RNNs. The basic structure of an LSTM includes an input layer, LSTM layer, hidden layers, and an output layer. Within the LSTM layer, each LSTM unit contains several key components: Cell State, Input Gate, Forget Gate, Output Gate, and Hidden State. When LSTMs are employed for time series data classification tasks, data is initially decomposed into individual time steps through the Input Layer and then fed into the LSTM layer. This layer captures long-term and short-term relationships within the time series data by maintaining, ignoring, or updating information through an internal state and three gate structures. The output from the LSTM layer is then passed to one or more Hidden Layers for further feature extraction, with the final classification result being produced by the output layer.

4. HYPERPARAMETER TUNING

Hyperparameter tuning plays a crucial role in the training of ML and deep learning models, as the choice of hyperparameters directly affects the performance, learning capability, and generalization ability of the model. During the training of various models mentioned before, such as RF, MLP, 1D CNNs, and LSTMs, experimenting with multiple combinations of hyperparameters is an effective method to find the relatively optimal hyperparameter settings.

4.1. RF Classifier Hyperparameter Tuning

The use of the grid search through "GridSearchCV" tool from the Scikit-learn python's library is proposed to systematically explore and optimize the hyperparameter settings of the RF classifier (Arrinda, Oyarbide, Macicior, & Muxika, 2021). Initially, we defined a search space containing various combinations of hyperparameters, including the number of decision trees, the maximum depth of the trees, the minimum number of samples required to split an internal node, the minimum number of samples required at a leaf node, and whether bootstrap sampling is used.

"GridSearchCV" tested each of the 2,400 different hyperparameter combinations defined in our search space and employed cross-validation to comprehensively assess the performance of each combination. The training dataset was divided into five subsets, with one subset being used as the validation set to evaluate the model and the remaining four subsets for training. The performance of each combination was assessed based on the average results of these five validations.

The optimal combination of hyperparameters for the model was finally identified by analyzing and comparing. After determining the best hyperparameters, these parameters were used with the full training dataset to conduct the final training of the random forest classifier, ensuring the model achieved optimal predictive performance.

4.2. Neural Networks Hyperparameter Tuning

The grid search method used for identifying the optimal hyperparameters of the RF Classifier was considered unsuitable for neural networks due to time cost concerns. Hence, the "Keras Tuner" Python's library is proposed to perform hyperparameter optimization through random search of neural network based models. Similar to "GridSearchCV", before starting the random search, a search space for each model is defined. However, the distinct feature of the random search method provided by "Keras Tuner" is that it does not attempt every possible combination of hyperparameters. Instead, it randomly selects n combinations of hyperparameters from the defined search space to experiment with. A key advantage of this approach is its ability to significantly reduce the search time while still maintaining the possibility of discovering well-performing hyperparameter sets.

5. RESULTS

The most suitable data processing approach and the most optimal hyperparameters for each model are shown in Figure 7. Each trained model has been validated both by virtual datasets and experimental datasets.

RandomForest
Data processing

- None

Hyperparameter

- Num of estimators: 500
- Bootstrap: True
- Max_depth: None
- Min_samples_leaf: 1
- Min_samples_split: 2

CNN
Data processing

- Sliding Window Size: 100
- Normalization Strategy: By Window
- Normalization Method: Z-Score

Hyperparameter

- Num of Module: 1
- Num of Convolutional Layer: 1
- Num of kernels: 96
- Kernel size: 5
- Activation function: relu
- L1 regularizer: No
- L2 regularizer: 0.00011731100894294
- MaxPooling: Yes
- Pool size: 3
- Dropout: No
- Num of Full Connected layer: 1
- Num of units: 256
- Activation function: relu
- L1 regularizer: 0.00020918054059229
- L2 regularizer: 0.00102808799175713
- Dropout: Yes
- Dropout rate: 0.3
- Learning rate: 0.0003967960974673

MLP
Data processing

- None

Hyperparameter

- Num of Hidden Layer: 1
- Dropout: Yes
- Dropout rate : 0.5
- Num of hidden units: 64
- Activation function: relu
- L1 Regularizer: 2.5099081541393e-05
- L2 Regularizer: 0.0004041274348191
- Learning rate: 0.000494137845717255

LSTM
Data processing

- Sliding Window Size: 30
- Normalization Strategy: By Feature
- Normalization Method: Min-Max

Hyperparameter

- Num of LSTM units: 256
- Dropout: No
- Recurrent dropout: Yes
- Recurrent dropout rate: 0.3
- L1 regularizer: No
- L2 regularizer: 2.5732638229762e-05
- Num of Hidden Layer: 1
- Num of Hidden units: 160
- Activation function: tanh
- L1 regularizer: 0.000423110271510091
- L2 regularizer: 8.380137024491e-05
- Dropout: No

Figure 7 The most suitable data processing approach and the most optimal hyperparameters for each model.

5.1. Validation with virtual dataset

The virtual dataset comprises one charging data set generated by a normal battery electric equivalent circuit model and other two charging data sets generated respectively by ISC battery electric equivalent circuit models, characterized by short-circuit resistances of 10Ω and 510Ω, respectively.

The validation results of models of the instantaneous feature-based method are shown in Figure 8, whereas the results for models utilizing historical feature-based methods are illustrated in Figure 9. These figures illustrate the temporal variation of cell voltage (Vcell) during the constant current charging process of normal batteries and batteries experiencing ISC faults with short circuit resistances of 10 ohms and 510 ohms. Furthermore, the figures depict the fault detection outcomes at each time point during the charging process, as predicted by the RF model and the RF model integrated with MLP. The prediction outcomes are marked in blue and red, indicating correct predictions and incorrect predictions, respectively.

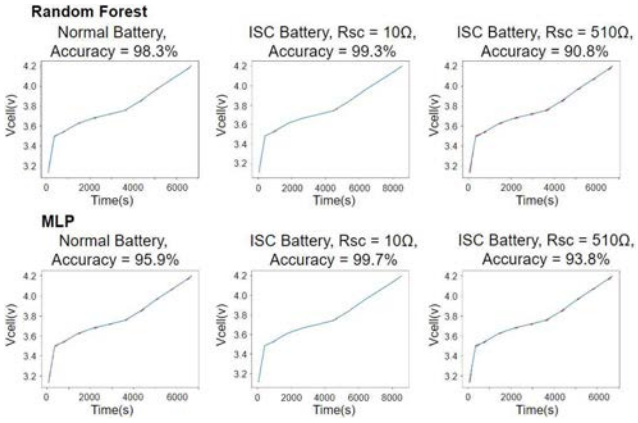


Figure 8: Validation Results of models of Instantaneous Feature-based method with virtual dataset. The blue line is the Vcell vs time, and the red points represent the moments where the prediction is wrong.

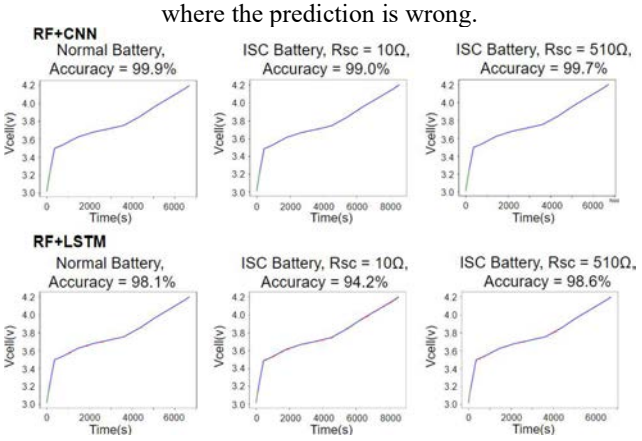


Figure 9: Validation Results of models of Historical Feature-based Methods with virtual dataset. The green part represents the moments when the charging amount is still less than 1% nominal capacity.

5.2. Validation with experimental dataset

The experimental dataset consists of two datasets. One of them is the charging dataset of a real battery without any fault. The other one is the charging dataset from the same battery connected with a 10Ω short-circuit resistance emulating an ISC condition.

The battery model used to generate the virtual data assumed a state of health (SoH) of 100%, whereas the battery utilized for the experimental dataset did not have the exact same SoH. Hence, Eq. (4) should be employed for calculating the voltage difference of the experimental dataset.

$$V_{diff\_\exp\_k} = (V_k - V_{k-1\%c}) \cdot SoH_{exp} \qquad (4)$$

Figure 10 and Figure 11 illustrate the validation results of Instantaneous Feature-based method models and Historical Feature-based method models respectively. These figures also use red and blue markers to denote the accuracy of predictions in relation to the actual conditions, where red indicates a mismatch between predicted and actual label, and blue signifies correct predictions.

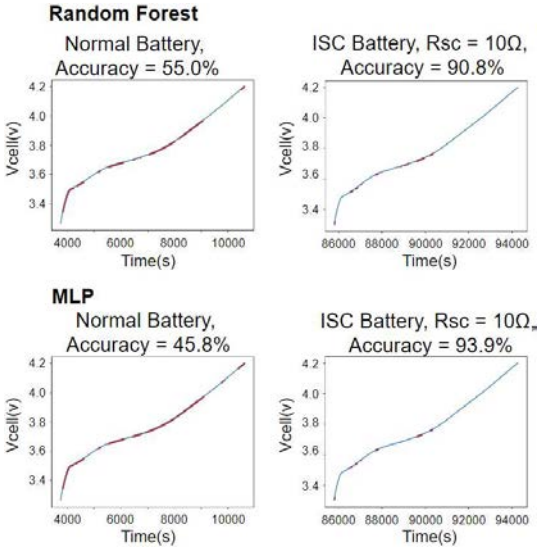


Figure 10: Validation Results of models of Instantaneous Feature-based method with experimental dataset.

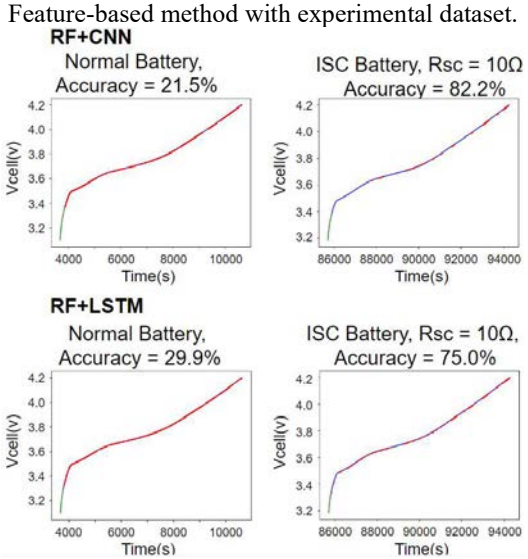


Figure 11: Validation Results of models of Historical Feature-based Methods with virtual dataset.

6. DISCUSSION

This study introduced four innovative methods for detecting ISC faults, all of which demonstrated a prediction accuracy above 90% (and 8 of 12 close to 99%) during the validation process with virtual data, including the identification of normal data and ISC data with 10Ω and 510Ω short-circuit resistances. The validation results are presented in Table1.

Table 1. Validation Accuracy of models on Different Validation Dataset.

Dataset	RF	RF+MLP	RF+CNN	RF+LSTM
Normal (Virtual)	98.3%	95.9%	99.9%	98.1%
Normal (Experimental)	55.0%	45.8%	21.5%	29.9%
10 Ω (Virtual)	99.3%	99.7%	99.0%	94.2%
10 Ω (Experimental)	90.8%	93.9%	82.2%	75.0%
510 Ω (Virtual)	90.8%	93.8%	99.7%	98.6%

The RF method, being the earliest and simplest instantaneous feature-based method, shows excellent performance in distinguishing between ISC and normal data within the virtual dataset. It achieved an accuracy of 98.3% for normal data and 99.3% for severe ISC detection (lower short-circuit resistance values), though its accuracy slightly decreased to 90.8% for early-stage ISC detection (higher short-circuit resistance values). To enhance the model's accuracy in predicting high resistance value ISC conditions, three other detection methods based on model stacking were proposed, using RF as the base model whose output serves as additional input features for other models.

The combination of RF with MLP slightly improved the accuracy for high resistance value ISC conditions to 93.8%, at the cost of reduced accuracy for normal battery data predictions, which fell to 95.9%. This is a consequence of the weight distribution of RF's estimators provided by the MLP.

The combination of RF with CNN performed best in virtual data validation. It achieved a 1.6% increase in accuracy for normal data predictions, reaching 99.9%. Moreover, this method improved the detection accuracy for 510Ω short-circuit resistance ISC from 90.8% to 99.7%. Nonetheless, its accuracy for detecting lower resistance ISC slightly fell by 0.3% to 99.0% compared to using RF alone.

The combination of RF with LSTM networks increased the prediction accuracy for high resistance ISC detection to 98.6% while maintaining the accuracy for normal data predictions at 98.1%. Nevertheless, lower resistance ISC detection slightly decreased to 94.2%.

The validation process with experimental data has shown a decrease in accuracy. All models experienced a significant performance decline in the validation with the experimental dataset, especially in predicting normal data. The instantaneous feature-based methods outperformed the historical feature-based methods overall in the experimental dataset. RF and RF+MLP maintained over 90% accuracy in predicting low resistance ISC conditions, but their accuracy in predicting normal data dropped to 55% and 45.8%, respectively. The HFM-based methods had less than 30% accuracy in predicting normal data, and their detection accuracy for ISC faults did not reach 80%.

This performance drop could be attributed to overfitting on virtual normal data, as the virtual normal battery could generate only a single set of normal battery charging data, and to discrepancies between data generated by equivalent circuit models and real data, preventing the RF model from accurately learning subtle changes in real conditions. The performance of model stacking methods was impacted by the base model RF's performance; if RF could not provide accurate predictions, the overall performance of the stacked models was negatively affected.

In real applications, as the performance gap between electric equivalent circuit model and actual battery increases, and therefore, the negative effect of stacked models will be amplified, making the RF model a preferred choice for ISC detection algorithms in the absence of real data. Nonetheless, if the performance gap could be resolved, or if real-life data could be used to further train the stacked models, models like RF+CNN could achieve significantly higher accuracy levels in early-stage ISC detection compared to the RF model alone.

7. CONCLUSION

This paper developed four methods for detecting ISC faults. Data is obtained from simulations and experiments at lab level. The ISC fault detection methods are trained using the virtual data. The normalization method, normalization strategy, are performed by using an exhaustive method and hyperparameter tuning is done by grid-search and random-search. After the training and hyperparameter tuning, these methods have been evaluated respectively by conducting validations and performance comparisons with both the virtual and experimental datasets.

The validation with virtual data shows that the historical feature-based method combining RF and CNN demonstrated superior performance. However, the limitations of virtual data became apparent during the validation with experimental data. The base model, the RF model, fails to achieve satisfactory prediction results on the experimental dataset. It suffers a drop of accuracy from 98.3% to 55% in describing data without ISC. This limitation further impacted the overall performance of the RF+CNN methods in the experimental data validation, having higher accuracy drops than the ones observed in RF.

To address this issue in future research, one of the potential approaches could be integrating digital twin and cloud computing technologies. Digital twins can facilitate the collection of extensive real-world data to refine the model, while cloud computing, combined with the gathered real-world data, can enable continuous learning for the model.

This method ensures the model's adaptability to real-world data. Moreover, as a battery's SoH gradually declines, affecting its charging data throughout its use, continuous learning can also allow the model to adjust to these changes. This synergy of multiple technologies considerably augments the flexibility and universality of ISC detection systems, equipping them to accommodate a wider array of scenarios and conditions.

ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon Europeresearch and innovation programme under grant agreement No 101103821. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

NOMENCLATURE

SoH_{exp}	*battery's SoH in the experimental dataset*
V_k	voltage at the kth moment
$V_{k\text{-}1\%c}$	voltage at the moment before charging 1% of the nominal capacity prior to the k^{th} moment.
V_{diff_k}	voltage difference of virtual dataset between k^{th} and k-1%c^{th} moment
$V_{diff\_exp\_k}$	voltage difference of experimental dataset between k^{th} and k-1%c^{th} moment
x	unnormalized original value
x_{min}	minimum value among dataset
x_{max}	maximum value among dataset
x_{norm}	normalized value
μ	mean value of dataset
σ	standard deviation of dataset

REFERENCES

Arrinda, M., Oyarbide, M., Macicior, H., & Muxika, E. (2021). Unified Evaluation Framework for Stochastic Algorithms Applied to Remaining Useful Life Prognosis Problems. *Batteries*, *7*(35), 1–27. https://doi.org/https://doi.org/10.3390/batteries7020035

Arrinda, M., Oyarbide, M., Macicior, H., Muxika, E., Popp, H., Jahn, M., Ganev, B., & Cendoya, I. (2021). Application Dependent End-of-Life Threshold Definition Methodology for Batteries in Electric Vehicles. *Batteries*, *7*(1), 12. https://doi.org/10.3390/batteries7010012

Chatterjee, S., Kumar Gatla, R., Sinha, P., Jena, C., Kundu, S., Panda, B., Nanda, L., & Pradhan, A. (2023). Fault detection of a Li-ion battery using SVM based machine learning and unscented Kalman filter. *Materials Today: Proceedings*, *74*. https://doi.org/10.1016/j.matpr.2022.10.279

Cheng, X., Li, X., & Ma, X. (2023). A method for battery fault diagnosis and early warning combining isolated forest algorithm and sliding window. *Energy Science and Engineering*, *11*(12), 4493–4504. https://doi.org/10.1002/ESE3.1593

Diouf, B., & Pode, R. (2015). Potential of lithium-ion batteries in renewable energy. In *Renewable Energy* (Vol. 76). https://doi.org/10.1016/j.renene.2014.11.058

naha, A., Khandelwal, A., Agarwal, S., tagade, piyush, Hariharan, K. S., Kaushik, A., Yadu, A., Mayya Kolake, S., Han, S., & oh, B. (2020). *internal short circuit detection in Li-ion batteries using supervised machine learning*. https://doi.org/10.1038/s41598-020-58021-7

Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *IARJSET*. https://doi.org/10.17148/iarjset.2015.2305

Ren, D., Feng, X., Liu, L., Hsu, H., Lu, L., Wang, L., He, X., & Ouyang, M. (2021). Investigating the relationship between internal short circuit and thermal runaway of lithium-ion batteries under thermal abuse condition. *Energy Storage Materials*, *34*. https://doi.org/10.1016/j.ensm.2020.10.020

Schmid, M., & Endisch, C. (2022). Online diagnosis of soft internal short circuits in series-connected battery packs using modified kernel principal component analysis. *Journal of Energy Storage*, *53*. https://doi.org/10.1016/j.est.2022.104815

Wang, H., Nie, J., He, Z., Gao, M., Song, W., & Dong, Z. (2023). A reconstruction-based model with transformer and long short-term memory for internal short circuit detection in battery packs. *Energy Reports*, *9*. https://doi.org/10.1016/j.egyr.2023.01.092

Xie, J., Zhang, L., Yao, T., & Li, Z. (2020). Quantitative diagnosis of internal short circuit for cylindrical li-ion batteries based on multiclass relevance vector machine. *Journal of Energy Storage*, *32*. https://doi.org/10.1016/j.est.2020.101957

Yang, N., Song, Z., Amini, M. R., & Hofmann, H. (2022). Internal Short Circuit Detection for Parallel-Connected Battery Cells Using Convolutional Neural Network. *Automotive Innovation*, *5*(2). https://doi.org/10.1007/s42154-022-00180-6

Zhan, J., Deng, Y., Ren, J., Gao, Y., Liu, Y., Rao, S., Li, W., & Gao, Z. (2023). Cell Design for Improving Low-Temperature Performance of Lithium-Ion Batteries for Electric Vehicles. In *Batteries* (Vol. 9, Issue 7). https://doi.org/10.3390/batteries9070373

Zhang, G., Wei, X., Tang, X., Zhu, J., Chen, S., & Dai, H. (2021). Internal short circuit mechanisms, experimental approaches and detection methods of lithium-ion batteries for electric vehicles: A review. In *Renewable and Sustainable Energy Reviews* (Vol. 141). https://doi.org/10.1016/j.rser.2021.110790

Zubi, G., Dufo-López, R., Carvalho, M., & Pasaoglu, G. (2018). The lithium-ion battery: State of the art and future perspectives. In *Renewable and Sustainable Energy Reviews* (Vol. 89). https://doi.org/10.1016/j.rser.2018.03.002

BIOGRAPHIES

Zihong Zhang received the B.S. degree in electrical engineering and automatization in 2019 at Shanghai Institute of Technology, Shanghai, China. He obtained M.S. in computational methods in sciences at University of Navarra, Pamplona, Navarra, Spain in 2021. In 2023, he completed the second M.S. in Artificial Intelligence at Tecnun, University of Navarra, Donostia, Gipuzkoa, Spain. His main responsibility in CIDETEC concentrates on development machine learning and deep learning models for applications in fault diagnosis, predictive maintenance, digital twins for lithium-ion batteries.

Mikel Arrinda received the B.S. degree in industrial electronic engineering in 2012 at MU, Mondragon, Basque Country, Spain. In 2013 completed his studies with a M.S. in integration of renewable energy sources into the electricity grid at EHU, Bilbao, Basque Country, Spain. In 2020, he obtained his Phd degree in apply engineering at CIDETEC institute for Energy Storage. His research interest is focused on lithium-ion batteries in terms of electric-thermal-aging modeling, SoX algorithms, thermal control strategies, digital twin applications, AI implementation and safety measures.

Jon Perez received the B.S. degree in electrical engineering in 2020 at Tecnun University of Navarre, Donostia – San Sebastian, Basque Country, Spain. In 2022, he received the M.Sc degree in Energy and Power Electronics from Mondragon University, Mondragon, Basque Country, Spain. From 2022 onwards, he is pursuing a PhD degree in applied engineering at CIDETEC institute for Energy Storage. His research interests are focused on lithium-ion batteries aging detection, SoX algorithms and early detection of Thermal Runaway and other failures.

Continuous Test-time Domain Adaptation for Efficient Fault Detection under Evolving Operating Conditions

Han Sun[1], Kevin Ammann[2], Stylianos Giannoulakis[3], and Olga Fink[4]

[1,4] *EPFL, Ecublens, Vaud, 1024, Switzerland*

han.sun@epfl.ch

olga.fink@epfl.ch

[2, 3] *Sulzer, Winterthur, Zürich, 8401, Switzerland*

Kevin.Ammann@sulzer.com

stylianos.giannoulakis@sulzer.com

ABSTRACT

Fault detection is crucial in industrial systems to prevent failures and optimize performance by distinguishing abnormal from normal operating conditions. Data-driven methods have been gaining popularity for fault detection tasks as the amount of condition monitoring data from complex industrial systems increases. Despite these advances, early fault detection remains a challenge under real-world scenarios. The high variability of operating conditions and environments makes it difficult to collect comprehensive training datasets that can represent all possible operating conditions, especially in the early stages of system operation. Furthermore, these variations often evolve over time, potentially leading to entirely new data distributions in the future that were previously unseen. These challenges prevent direct knowledge transfer across different units and over time, leading to the distribution gap between training and testing data and inducing performance degradation of those methods in real-world scenarios. To overcome this, our work introduces a novel approach for continuous test-time domain adaptation. This enables early-stage robust anomaly detection by addressing domain shifts and limited data representativeness issues. We propose a Test-time domain Adaptation Anomaly Detection (TAAD) framework that separates input variables into system parameters and measurements, employing two domain adaptation modules to independently adapt to each input category. This method allows for effective adaptation to evolving operating conditions and is particularly beneficial in systems with scarce data. Our approach, tested on a real-world pump monitoring dataset, shows significant improvements over existing domain adaptation methods in fault detection, demonstrating enhanced accuracy and reliability.

1. INTRODUCTION

Fault detection aims to identify evolving faults or degradation in complex industrial systems, aiming to prevent system failures or malfunctions. Early and robust fault detection is essential for optimizing equipment performance and minimizing maintenance and unavailability costs. Recently, data-driven methods have been widely applied to fault detection facilitated by the growing availability of system condition monitoring data (Fink et al., 2020). However, these methods often assume the availability of abundant, representative training datasets to learn a data distribution that is applicable across all relevant operating and environmental conditions. Such representative training datasets are frequently not available due to the high diversity of systems and the wide range of operating conditions. This issue is particularly acute for newly installed or refurbished units with limited observation periods. A potential solution to this problem is to transfer knowledge and operational experience from fleet units with extensive and relevant data to those lacking representative training data. This approach leverages the rich experience and datasets of 'experienced' units to enhance the learning and performance of less experienced ones, aiming to bridge the gap in data availability and representativeness across the fleet. However, such knowledge transfer might lead to insufficient performance, as data-driven methods typically assume

identical and independent distributions (i.i.d) between training and testing data, which does not hold true in real-world industrial complex systems with varying operating conditions and dynamic environments. This leads to significant discrepancies in data distribution between fleet units. Consequently, a model trained on one unit may perform poorly when applied to another, evidenced by a high rate of false alarms, preventing them from benefiting from the existing fleet knowledge.

A substantial amount of research has been performed to address such a challenge by applying domain adaptation (DA) approaches (Yan et al., 2024), which aim to bridge the domain shift between a labeled source and a related unlabeled target domain. However, The scarcity of faulty data in industrial systems introduces specific challenges for DA in fault detection because industrial applications typically lack labeled source data for supervised learning. Furthermore, these methods usually assume discrete source and target domains. However, the operating conditions of complex systems evolve over time, leading to continuous domain shifts within the same unit. Therefore, domain adaptation should not only occur between units but also be continuously applied within a unit, rather than assuming a single, discrete target domain, to ensure robust fault detection.

In this work, we propose a novel approach for fleet-wide continual test-time domain adaptation, aiming to achieve robust anomaly detection across different units of a fleet over time. Our proposed fault detection framework, based on signal reconstruction, integrates a domain adaptive module specifically designed to address the dynamic and evolving environments of complex industrial systems. This approach aims to enhance robustness and adaptability in fault detection within these challenging contexts. To prevent overfitting to the faulty data distribution during adaptation, we categorize the input variables into two groups: control parameters and sensor measurements. We then integrate two domain adaptive modules to adapt to the data distribution of each category separately. This strategy enables us to distinguish between normal variations inherent within the systems and abnormal changes in operating status, thereby improving the accuracy of our anomaly detection framework. By integrating adaptation into the basic fault detection pipeline, TAAD facilitates the transfer of operational experience between different units of a fleet, thereby benefiting from the collective knowledge of the fleet. TAAD has been evaluated on a real-world pump monitoring dataset, and the results demonstrate notable improvements compared to other domain adaptation methods. Our proposed framework is transferable to other industrial applications and enables more timely and robust fault detection in complex industrial systems.

2. Related Work

2.1. Fault Detection in Prognostics and Health Management

Prognostics and Health Management (PHM) seeks to enhance equipment performance and minimize costs by enabling precise detection, diagnosis, and prediction of the remaining useful lifetime as accurately as possible. It integrates the detection of an incipient fault (fault detection), its isolation, the identification of its origin, and the specific fault type (fault diagnostics), along with the prediction of the remaining useful life (Fink et al., 2020). Fault detection aims to identify faulty system conditions based on current operating conditions and gathered condition monitoring data. The complexity of real-world industrial systems poses specific challenges for achieving accurate and robust fault detection. First, faulty data is scarce in real industrial systems. Failures in critical systems, such as power or railway systems, are infrequent. Furthermore, it often takes a considerable amount of time for a system to degrade to the point of failure or the end of life. As a result, faults are often never or seldom encountered during limited time periods and are therefore absent in training datasets.

Consequently, one of the main research directions in fault detection has focused on unsupervised learning, which can be categorized into three main directions(Ruff et al., 2021). Probabilistic models aim to approximate the normal data probability distribution. The estimated distribution mapping function can then be used as an anomaly score. Different deep statistical models have been applied for probability-based anomaly detection, such as energy-based models (EBMs) (Zhai, Cheng, Lu, & Zhang, 2016). One-class classification models directly learn a discriminative decision boundary that corresponds to a desired density level of normal samples, instead of estimating the full density (Ruff et al., 2021). This approach aims to learn a compact boundary that encloses the normal data distribution (J. Wang, Qiu, Liu, Yu, & Zhao, 2018; Z. Zhang & Deng, 2021). Reconstruction-based methods learn a model, such as autoencoders (AEs), are optimized to reconstruct the normal data samples well and detect via reconstruction error (Lai et al., 2023; Hu, Zhao, & Peng, 2022). These models are expected to fit the data distribution under healthy conditions and then raise an alarm for predictions with large deviations when the test data distribution is significantly different from the learned distribution. Thus, the reconstruction error serves as an anomaly score for detecting faults.

Other studies have focused on semi-supervised learning, where it is presumed that a limited number of faulty data samples are accessible for training (Ramírez-Sanz, Maestro-Prieto, Arnaiz-González, & Bustillo, 2023).

2.2. Fleet Approaches for Fault Detection

Unsupervised fault detection, as discussed in section 2.1, relies on the assumption that all possible normal conditions of the system can be learned from a sufficiently large and representative training dataset. However, collecting a dataset representative enough for new systems or refurbished units to cover all possible normal operating conditions within a short time period is unlikely. While extending the observation period can facilitate the collection of more comprehensive data, it also hinders early monitoring of the system. In such cases, transferring operational experience from other similar units with longer and more representative data can significantly enhance robust detection at an early stage. These units can be grouped into a fleet, where each unit shares similar characteristics (Leone, Cristaldi, & Turrin, 2017). A good example would be a fleet of gas turbines or cars produced by the same manufacturer, albeit with different system configurations, operating under varying conditions in different parts of the world (Fink et al., 2020).

The direct transfer of fleet knowledge assumes identical and independent distributions (i.i.d) between training and testing units. However, this assumption often does not hold for complex industrial systems, which are characterized by varying operating conditions and changing environments. This discrepancy poses a significant challenge in transferring a developed model across different units within the fleet. Traditional methods focus on identifying units that are similar enough to form sub-fleets (Leone, Cristaldi, & Turrin, 2016; Liu, Tan, Zhen, Yin, & Cai, 2018; Michau, Palmé, & Fink, 2018; Michau & Fink, 2019). Such methods depend on the entire fleet sharing sufficient similarity and fail when units under homogeneous conditions do not exist or cannot be identified. Recently, domain adaptation has been used to transfer knowledge between units or between different operating conditions within the same unit (Yan et al., 2024), a topic discussed in section 2.3.

2.3. Domain Adaptation Applied to Fault Detection

A substantial amount of research in the field of PHM has focused on domain adaptation, a subtopic of transfer learning, including discrepancy-based methods (J. Zhang et al., 2022; Qian, Wang, Zhang, & Qin, 2023) and adversarial-based methods (Michau & Fink, 2021; Qian, Qin, Luo, Wang, & Wu, 2023; Nejjar, Geissmann, Zhao, Taal, & Fink, 2024). These DA methods aim to align the data distribution between the source and target domains, assuming that the target samples available are abundant enough to represent target data distribution. However, this assumption does not hold true for newly installed systems with limited data samples collected, which prevents prompt system monitoring as discussed above (Michau & Fink, 2021). Furthermore, these methods typically assume one or more discrete, static target domains and attempt to adapt to them. However, operating conditions often evolve continuously over time, potentially leading to unseen distribution shifts in the future. Therefore, it is necessary to continuously adapt to domain shifts on the fly, rather than assuming a single discrete target domain (Q. Wang, Fink, Van Gool, & Dai, 2022).

Test-time adaptation (TTA) aims to adapt a source-pretrained model to a target domain without using any source data. The model is dynamically updated on the fly, based on the current data batch, without exposure to the entire target data set. Representative methods utilize batch normalization, estimating and normalizing mean and variance on each batch to update the model (D. Wang, Shelhamer, Liu, Olshausen, & Darrell, 2021; Liang, Hu, & Feng, 2020). Thus, TTA can be applied to adapt batch data online, accommodating continuous domain shifts for fault diagnosis (Q. Wang, Michau, & Fink, 2019). Although this branch of methods can be directly adapted for the fault diagnostic task, it is not suitable for unsupervised fault detection. In scenarios of unsupervised fault detection, where detection is based on deviation from the norm, applying TTA to the current batch of data with unknown labels may cause the model to unintentionally fit potentially faulty data within this batch. Consequently, the anomalies might not be recognized as out-of-distribution, leading to a reduced ability of the model to identify faults based on prediction errors.

To conclude, robust fault detection in PHM at the early stage encounters the challenge of data scarcity. Fleet approaches help units that are newly taken into operation benefit from fleet knowledge, while their transferability is constrained by the high variability of system operating conditions within the fleet. Current DA methods applied for fault detection cannot simultaneously address all of the challenges we discussed above. They either fail to adapt to continuous domain shifts or are incompatible with the limited data and label availability elaborated above.

3. METHODOLOGY

3.1. Problem Definition

The primary motivation of this research is to transfer knowledge from one system, which has abundant monitoring data, to other systems or fleets operating under varying conditions. Often, these systems and fleets are newly taken into operation, for which only a limited amount of observations can be collected for training. Their data distribution can evolve continuously due to changes in operating conditions and environmental factors. The objective is to adapt the prediction model trained on the original system, enabling it to make accurate predictions for new systems and fleets, even when only a few training samples are available. Given:

- abundant healthy training data from the source system:

$X_s = [x_1^s, \cdots, x_n^s]$, where s denotes the source domain and n denotes the number of data samples from the source domain, and

- limited observed normal data from the target domain: $X_t = [x_1^t, \cdots, x_m^t]$, where t denotes the target domain and m denotes the number of available data samples from the target domain,

the goal here is to achieve robust fault detection in the target domain t.

The proposed method takes into account limited data availability and varying operating conditions, specifically addressing scenarios where: 1) no anomalies are available for training; 2) only limited target data is available for adaptation; and 3) continuous changes in operating conditions occur during test time.

3.2. Reconstruction-based Anomaly Detection Framework

We develop a reconstruction-based anomaly detection pipeline, which achieves robust fault detection by continuously adapting to novel operating conditions, as depicted in Figure 1. This approach utilizes an autoencoder (AE), denoted as f_θ, trained exclusively on normal source data samples, X_s, for the purpose of signal reconstruction. The goal of f_θ is to accurately model the normal data distribution of X_s with accurate predicted signal value $\hat{X_s}$. The training objective is to minimize the mean-squared error (MSE) between the original data samples X_s, and their reconstructed counterparts $\hat{X_s}$:

$$loss_{MSE} = \frac{1}{n}\sum_{1}^{n}\left(X_s - \hat{X_s}\right)^2 \tag{1}$$

where n denotes the number of training samples. Thus, on the healthy source dataset, we expect a small residual value $\hat{r_s} = \hat{X_s} - X_s$. During testing, data samples generating large residuals are considered out-of-distribution and subsequently labeled as anomalies.

The autoencoder architecture consists of two parts: an encoder f_e and a decoder f_d. f_e comprises three fully-connected layers each followed by a batch normalization layer and a ReLU activation function, which consecutively map the original signal input to feature dimensions of 50, 50, and 10. f_d follows a similar architecture but without batch normalization layers, decoding the latent representation from 10 to 50, 50, and then back to the original signal dimension.

3.3. Anomaly Score and Anomaly Detection

During test time, we compute the fault label $y \in [0, 1]$ based on the reconstruction result. 0 denotes a healthy sample while 1 indicates a faulty sample. Given the i_{th} data sample $X_i = [x_i^1, ...x_i^k]$, we compute its relative residual:

$$r_i = \frac{|\hat{X_i} - X_i|}{\bar{X}_{t\_training}} \tag{2}$$

given its predicted reconstruction result $\hat{X_i}$. k indicates the input dimension. $\bar{X}_{t\_training}$ represents the mean value of target data samples for training (including validation data), which helps scale the residual values. The anomaly score s_i is calculated by integrating the scaled residual values across all sensors:

$$s_i = \frac{1}{k}\sum_{j=1}^{k} r_i^j + max\sum_{j=1}^{k} r_i^j \tag{3}$$

To avoid false detection by outliers with extremely large residuals, the computed anomaly score is smoothed within a certain window length l:

$$s_{i\_smooth} = min\sum_{q=0}^{l-1} s_{i+q} \tag{4}$$

Anomalies are then detected based on s_{i_smooth}, using a threshold determined via statistical analysis of the healthy validation set. We identify the data sample X_i as an anomaly if:

$$s_{i\_smooth} > \alpha * \bar{r}_{t\_training} \tag{5}$$

where α is set empirically with a trade-off between the reduction of false alarms and sensitivity to faults.

In this case study, potential faults are reported and examined daily. Thus, the evaluation of abnormal conditions is conducted on a daily basis, where we compute the number of cumulative abnormal data samples of the day.

3.4. System Variables

Directly applying domain adaptation to the current anomaly detection framework can potentially cause the model to fit unknown abnormal samples in the target domain's current batch during test time, thus impairing the model's ability to detect those faults. To distinguish between data distribution shifts due to changing operating conditions and occurrences of abnormal operating status, we split the input parameters into two groups: $X = [x, w]$. w denotes control variables, indicating variables that control system conditions. These variables are set by the operators or by the control system to optimize the performance under specified conditions. x represents sensor measurements, which are sensor signals monitoring system components and reflecting real-time system states. Here, we assume that changes in the distribution of control variables do not necessarily indicate an abnormal status but rather distinct operating conditions to which we should adapt.

3.5. Test-time Domain Adaptation Anomaly Detection

Figure 2 illustrates our proposed cross-domain, reconstruction-based anomaly detection framework TAAD, inspired by recent advancements in test-time domain adaptation. This framework enables us to achieve robust anomaly detection across different domains through online adaptation. Based on the pretrained reconstruction framework introduced in 3.2, we integrate an adaptive module, h_ϕ, for test-time domain adaptation to bridge the domain gap between the source and target domains. The decision to incorporate a separate adaptive module, h_ϕ, rather than embedding adaptive layers directly into the reconstruction model, stems from limitations observed in unsupervised anomaly detection. TTA methods, such as AdaBN, adapt to each batch during test time, inevitably fitting the distribution of abnormal data points. This impairs the model's ability to distinguish between normal and abnormal samples. Instead, our adaptive module takes the predicted value and original controlled system variables as inputs. By excluding monitoring data signals from adaptation, this approach prevents overfitting to the potentially faulty data distribution, thereby preserving the model's capability to accurately distinguish anomalies.

The adaptive module is a simple network composed of two fully connected layers that map the group of control variables from its original feature dimension to 10, and then back to its original dimension; the first is followed by a batch normalization layer and ReLU activation, while the second is followed by ReLU activation only. This adaptive module exclusively processes the control variables w as its input. During the adaptation phase, the pre-trained autoencoder f_θ is frozen, and the adaptive module h_ϕ is trained on a few target data samples to predict Δx, aimed at compensating for the large prediction errors due to the domain gap between source and target data. To address continuous domain shifts during test time, an AdaBN layer is incorporated into the adaptive module. This layer updates its mean and variance based on batch statistics during test time. The predicted δx is then added to the original prediction made by f_θ to compensate for inaccurate predictions caused by operating condition domain shift.

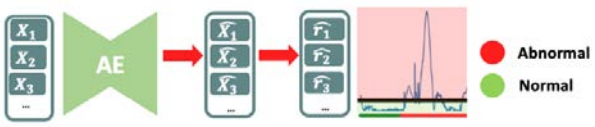


Figure 1. General pipeline of reconstruction-based unsupervised anomaly detection

Station	Pump	Seal Type	Operator Maintenance Reports
A	A-A	Type 1	primary& secondary seal replacement, primary& secondary seal leakage
	A-C	Type 2	secondary seal replacement, primary & secondary seal leakage
B	B-B	Type 1	seal replacement
	B-C	Type 2	secondary seal replacement, primary seal leakage
	B-D	Type 1	N/A

Table 1. Details on industrial pump dataset.

4. Case Study on Real-world Pump Dataset

4.1. Industrial Pump Dataset

In this case study, we aim to achieve early and robust fault detection while reducing false alarms under normal operating conditions. We evaluate our proposed method on a case study of a real industrial dataset, highlighting its effectiveness in achieving early fault detection and minimizing false positive alarm rates. The experiments are conducted on a real-world pump dataset, which comprises condition monitoring data collected from users with installations of various types of pumps in different locations. As a result, these data represent real-world, noisy data distributions and encompass a wide range of diverse domains, including operating conditions, environments, and pump types. In this dataset, we apply the proposed methodology to obtain robust adaptation across pumps, stations, and different pump types.

The selected dataset has two installation stations, with four Heat Transfer Fluid pumps installed at each, and the pumps are equipped with dual seals of two different types. Several seal failures were recorded for seven out of eight pumps during the data collection period. This dataset is marked by continuous changes in operating conditions as the controlled parameters are adapted by the operators regularly. We chose five pumps from this dataset with enough recorded data samples for validation for the case study, as summarized in Table 1. To comply with data policy requirements, we use virtual dates (year.month.day) to represent timelines. Fault durations reported by on-site operators may lack precision due to delayed inspections. Additionally, faults often occur significantly earlier than actual failures, and systems do not immediately return to normal conditions after maintenance of those failures. Given these factors, we consider data samples within a two-month window of any reported faults as uncertain regarding their health status. Therefore, we exclude them from both training data and the subsequent evaluation of false alarms.

We categorize the input variables according to the sub-components associated with the pumps of this dataset. The studied pumps are composed of four main parts: pump bearings, pump driver axial bearings, motor bearings, and seals. Since only seal faults are reported in this dataset, we specifically focus on performance and seal-related variables, while disregarding other

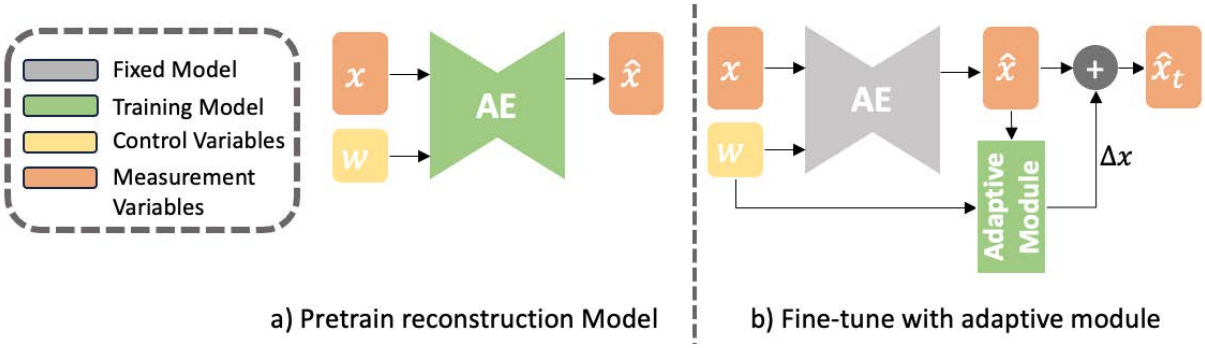


Figure 2. Test-time domain Adaptation Anomaly Detection (TAAD). a) We first pretrain the reconstruction-based anomaly detection model on the source dataset; b) For domain adaptative prediction, we add domain adaptive module and train it on the target training data.

parameter groups. The used parameters are displayed in table 2. Sensor monitoring data are available for both the drive end (DE) side and non-drive end (NDE) side of the pumps, reflecting the state of the dual seal on either side and its primary and secondary components. We exclude pump seal level variables due to their frequent manual adjustments, which do not accurately reflect the operating status of the pump seal. Instead, we focus on variables within the pump performance group, which represent the general operating condition of the current pump. In all our experiments, we designate the seal DE and NDE variables as x and the pump performance group variables as w inputs, as introduced in Section 3.4.

Group	Description	Variable
Pump DE Seal	Dual Seal Parameters	pump DE seal pressure
		pump DE seal pressure secondary
		pump DE seal temperature
		pump DE seal temperature secondary
		pump DE seal level
		pump DE seal level secondary
Pump NDE Seal	Dual Seal Parameters	pump NDE seal pressure
		pump NDE seal pressure secondary
		pump NDE seal temperature
		pump NDE seal temperature secondary
		pump NDE seal level
		pump NDE seal level secondary
Pump Performance	Pump Operating Conditions	pump uncorrected flow
		pump speed
		pump pressure suction
		pump pressure case
		pump uncorrected head
		pump uncorrected shaft power

Table 2. Groups of input variables of the industrial pump dataset

4.2. Details on Data Selection and Implementation

Robust fault detection is critical for newly established industrial systems, posing a challenge due to the brief operational history of such pumps. Our goal is to enable early fault detection capabilities with minimal data. To address this, we adopt a strategy where a source model is pretrained on a well-established pump with abundant data samples, then adapted to target pumps with limited operational data. This approach aims to achieve robust fault detection on target pumps despite the limited available data for these new installations.

We select pump B-C as our source domain, ensuring that abundant normal data samples are available for training. We train the source model on this pump using data samples from a 12-months period and validate it with an additional 12 months of healthy data. The remaining pumps are treated as target domains, each with limited training samples. For each pump, we use three months of normal data samples for training and one month for validation. The training and testing data for the target domains are summarized in Table 3. Due to sensor failures, some data samples lack measurements related to seal components. Considering that only four parameters are available for each seal, with each being crucial and independent of the other parameters, we exclude any data samples with missing measurements. We use the Min-Max Scaler to scale the signals to a range between 0 and 1.

During test time, we combine the 3-month training and 1-month validation data in the target domain to compute the threshold $\bar{r}_{t\_training}$ for anomaly detection. We set $\alpha = 1.5$ for domain adaptation within the same installation station and $\alpha = 2.0$ for adaptation across stations, as the domain gap is comparatively larger.

4.3. Evaluation Metrics for Robust and Early Detection

Given the real-world nature of our dataset, which features imbalanced data, limited collected samples, and uncertain labels for validation, traditional evaluation metrics such as F1 score

and accuracy may not be applicable. To align with needs and interests in real industrial application scenarios, we evaluate TAAD from two perspectives: 1) minimizing false alarms on normal data samples caused by domain shift and 2) achieving early detection of significant system faults.

Unsupervised fault detection relies on the assumption that the model learns the healthy data distribution and identifies deviating distributions as faults. As the operating conditions of complex industrial systems vary, novel operating conditions that have not been seen by the model are often identified as faults. Such false positive (FP) predictions need to be avoided. To assess the effectiveness of TAAD in reducing false alarms due to domain shift, we test TAAD on data collected under unseen healthy conditions with only positive samples. Thus, any reported faults are FP. We identify periods of known normal operations and evaluate the prediction results for these periods, excluding data samples from two months before and after any reported fault to avoid periods of potential pre-fault conditions. The evaluation periods vary for each pump due to data sample availability limitations, as detailed in Table 3. We determine the count of FP, which indicates inaccurately predicted faults, and compute the false positive rate: $\frac{FP}{FP+TP}$ (TP indicates true positives). The lower the rate, the better our adaption to novel operating conditions and our ability to avoid false alarms.

Faulty conditions vary in severity levels. The system or a specific component can continue to operate despite the occurrence of faults, gradually degrading until a complete failure. This gradual degradation can lead to more severe faults, ultimately stopping operation and potentially leading to secondary damages. Therefore, our goal is to achieve early detection before the faults are observed and recorded. For early detection of system faults, we summarize all reported faults for each pump in Table 3. We conduct an evaluation starting 14 days before the recorded fault date to determine the earliest point of detection achievable by TAAD. This evaluation, performed daily as stated in section 3.3, focuses on the first predicted abnormal day. This value indicates how early we can detect potential faults in the system and preemptively address them. Additionally, we report the number of days detected as anomalies within this 14-day window to assess detection robustness. A higher count of abnormal days following the initial fault observation indicates better robust detection capability in consistently issuing alarms, enhancing certainty to involve on-site inspection and minimizing missed faults.

5. Experimental Results Pump Case Study

Experiments on fault detection in this pump system, which includes two installation stations, involve two distinct case studies: intra-station transfer, where the domain gap is relatively smaller, and inter-station transfer, which has a larger domain gap. The proposed method is compared with AdaBN and MMD, as well as with the baseline model without adaptation on the target data. Performances are reported based on the evaluation metrics introduced in section ref. General experiment results are summarized in Table 4.

5.1. Case 1: Transfer within Station

In the first case study, we evaluate the adaptation performance of our TAAD applied to two pumps characterized by a relatively small domain gap. These two pumps are installed at the same station as pump B-C, on which we train the source model.

Pump B-D: Two seal leakages were observed after the training time period of this target pump. We visualize this case in Figure 3 for a better understanding. The yellow vertical lines mark the recorded first occurrence data of the faults. We scatter the predicted faults of each method on a daily basis. In the case of the earlier secondary seal leakage, occurring 4 months after the adaptation, all methods managed to detect it, however, TAAD not only detected it but also did so earlier and the detection is more robust with more true positives. Regarding the later primary seal leakage, which occurred 1 year 7 months after adaptation, all other methods failed to trigger any alarms as the operating condition evolved. In contrast, TAAD successfully detected the reported fault 9 days in advance, registering 8 abnormal days, thus demonstrating its superior adaptability to long-term changes in operating conditions. As AdaBN does not consider source data and thus tends to overfit current batch statistics, it fails to detect any fault. Furthermore, compared to MMD, TAAD effectively reduced false alarms, highlighting the robustness of our proposed approach, which benefits from avoiding overfitting to unstable and unrepresentative measurement variables during adaptation.

Pump B-B: A leakage in the secondary seal is reported 5 months after the training period for adaptation. Given the proximity of the event and the small domain gap, all methods were capable of predicting the fault 14 days in advance. In this scenario, TAAD significantly reduced the false alarm rate from 0.15 to 0.02 by adapting to changing operating conditions. Even when compared to MMD – which benefits from access to the source domain's training data for direct data distribution alignment – TAAD demonstrated superior adaptation capabilities by avoiding overfitting to the system's atypical operating status.

5.2. Case 2: Transfer across Stations

This case study involves transferring a model pre-trained on pump B-C from station B to station A, anticipating a significantly larger domain gap due to variations in environments and operational regimes across the stations.

Pump A-A: Two seal leakage faults were reported succes-

target domain	training time period on target domain	normal test time period	fault type	faulty time period
B-B	00.01.01-05.01	00.05.01-06.01	secondary seal replacement	00.10.07-10.14
B-D	00.01.01-04.01	00.11.01-01.01.01	secondary seal leakage	00.08.07-08.22
			primary seal leakage	01.11.12-11.13
A-A	00.01.01-00.05.01	00.06.01-00.07.01	secondary seal leakage	00.08.13-08.15
			primary seal leakage	00.10.14-11.17
A-C	00.01.01-00.05.01	01.07.01-01.08.01	primary seal leakage	00.07.10-08.10
			secondary seal leakage	01.12.26

Table 3. Experimental settings on pump dataset. We report the training time period for domain adaptation for each pump and the normal test time period with no observed faults in between for evaluating the false alarm rate. We report the fault type and the observed faulty time period of each occurrence of fault to evaluate the detection days in advance and the number of detected abnormal days within 14 days prior to the first reported date of the fault. The dates are represented by virtual dates with the same duration and intervals as in the real dataset due to the data privacy policy.

target domain	num of normal samples	false alarm rate ↓				fault type	detection days in advance ↑				num of detected abnormal days within 14 days ↑			
		baseline	AdaBN	MMD	TAAD		baseline	AdaBN	MMD	TAAD	baseline	AdaBN	MMD	TAAD
B-B	1903	0.15	0.03	0.05	**0.02**	secondary seal replacement	**14**	**14**	**14**	**14**	**12**	**12**	**12**	**12**
B-D	2275	**0**	**0**	0.01	**0**	secondary seal leakage	2	0	2	**4**	1	0	1	**3**
						primary seal leakage	0	0	0	**9**	0	0	0	**8**
A-A	1969	**0**	**0**	0.01	0.01	secondary seal leakage	1	0	10	**11**	1	0	2	**4**
						primary seal leakage	0	0	0	**1**	0	0	0	**1**
A-C	2135	**0**	**0**	0	0.07	primary seal leakage	0	0	0	**1**	0	0	0	**2**
						secondary seal leakage	0	0	0	**14**	0	0	0	**6**

Table 4. Experimental results on pump dataset. We compare our method with the baseline model w/o domain adaptation, AdaBN, and MMD, and report the false alarm rate and detection days in advance on 4 pumps across two stations. The best results are in bold.

sively 3 months after the training time period for adaptation on this target pump. For the first secondary seal leakage fault, TAAD achieved the earliest detection with the highest number of days detected as anomalies within this 14-day window before the recorded fault and only 1% false alarms during the normal operating period. Detecting the subsequent primary seal leakage proved much more difficult, as it occurred shortly after the maintenance of the previous fault and operated under unstable and significantly different operating conditions. Here, while all the compared methods failed to detect the fault, our proposed method managed to raise an alarm on the last day before the fault was reported.

Pump A-C: Two seal leakage faults were recorded. The other methods failed to detect either fault, whereas TAAD successfully reported both. However, in this challenging scenario, TAAD increased the false alarm rate by 7% due to the less accurate adaptation.

Generally, experimental results confirm that detecting faults is more challenging than in intra-station cases. Nonetheless, TAAD successfully detects faults under these challenging scenarios, outperforming other methods.

5.3. Discussion

We demonstrate the effectiveness of TAAD to achieve early and robust fault detections in the above two case studies. First, TAAD significantly reduces the false alarm rate under easy-to-detect scenarios compared to other methods, as proved in case 1 on pump B-B, when the target pump is installed within the same station with a smaller domain gap and the fault happens shortly after the adaptation training. Second, TAAD remains effective and robust long after the initial adaptation phase, as shown in the inboard seal leakage of pump B-D and the outboard seal leakage of pump A-C. Those cases demonstrate the ability of the proposed method to adapt to dynamic evolving operating conditions Third, TAAD achieves robust detection under significant domain shifts across different installation stations compared to other methods which fail to detect before the occurrence of faults, as shown in the experiments on pumps A-A and A-C.

In general, our TAAD achieves overall better performance than the other methods achieving earlier detections, and providing more continuous and robust detection within the time window before fault occurrences, all while maintaining a low false alarm rate.

6. Conclusions

In this paper, we propose an effective continuous test-time domain adaptation approach TAAD for efficient and robust anomaly detection under evolving operating conditions. This approach does not require labeled faulty data and needs only a minimal amount of normal data samples for adaptation. Such requirements align well with the practical needs of real-world

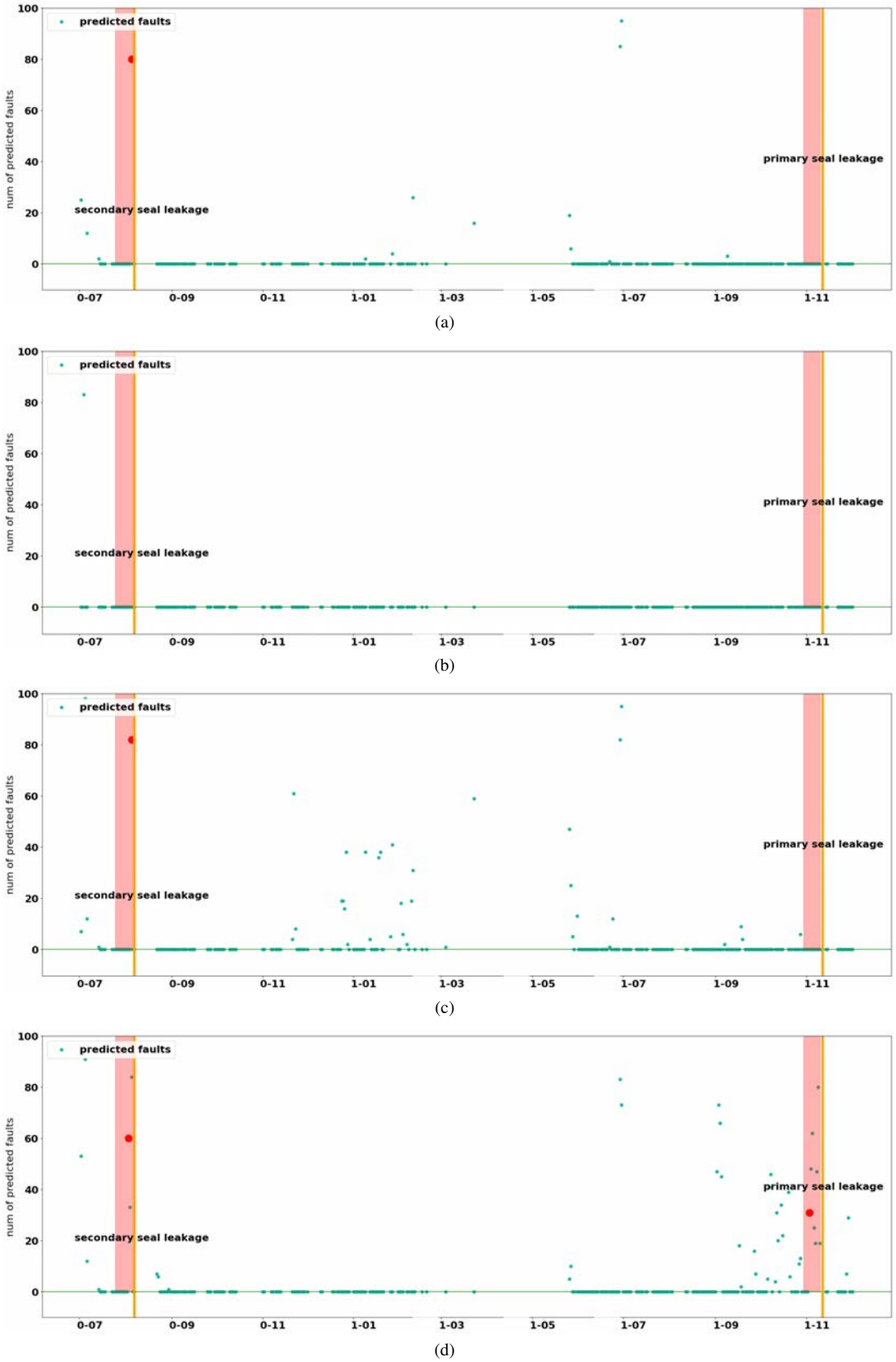


Figure 3. Comparison of performance of (a) baseline (b) AdaBN (c) MMD (d) TAAD for fault detection on Pump B-D. The number of predicted faults per day is plotted. The yellow lines mark the starting date of the reported faults. Red regions mark the 14-day time window before the occurrence of reported faults, and the red dots within this region mark the first detected abnormal day within this period.

industrial systems. We compared our method with two other representative domain adaptation methods. The experimental results demonstrate TAAD's effectiveness in achieving early fault detection under significant domain shifts, both across different stations and over time, while maintaining a low false alarm rate.

Despite its satisfying performance, we see potential improvements in the current method. First, our adaptive module continuously adapts to the current batch without considering the size of the domain gap. We hypothesize that the performance could be enhanced by re-training this module once a significant domain shift is detected. Second, the thresholding parameter α is currently determined empirically. An automatic adjustment of this parameter, taking into account both distribution shifts and operational requirements, could optimize the trade-off between minimizing false alarms and attaining prompt fault detection.

REFERENCES

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Hu, Z., Zhao, H., & Peng, J. (2022). Low-rank reconstruction-based autoencoder for robust fault detection. *Control Engineering Practice*, *123*, 105156.

Lai, K.-H., Wang, L., Chen, H., Zhou, K., Wang, F., Yang, H., & Hu, X. (2023). Context-aware domain adaptation for time series anomaly detection. In *Proceedings of the 2023 siam international conference on data mining (sdm)* (pp. 676–684).

Leone, G., Cristaldi, L., & Turrin, S. (2016). A data-driven prognostic approach based on sub-fleet knowledge extraction. In *14th imeko tc10 workshop on technical diagnostics: New perspectives in measurements, tools and techniques for systems reliability, maintainability and safety* (pp. 417–422).

Leone, G., Cristaldi, L., & Turrin, S. (2017). A data-driven prognostic approach based on statistical similarity: An application to industrial circuit breakers. *Measurement*, *108*, 163–170.

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning* (pp. 6028–6039).

Liu, L., Tan, E., Zhen, Y., Yin, X. J., & Cai, Z. Q. (2018). Ai-facilitated coating corrosion assessment system for productivity enhancement. In *2018 13th ieee conference on industrial electronics and applications (iciea)* (pp. 606–610).

Michau, G., & Fink, O. (2019). Unsupervised fault detection in varying operating conditions. In *2019 ieee international conference on prognostics and health management (icphm)* (pp. 1–10).

Michau, G., & Fink, O. (2021). Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. *Knowledge-Based Systems*, *216*, 106816.

Michau, G., Palmé, T., & Fink, O. (2018). Fleet phm for critical systems: bi-level deep learning approach for fault detection. In *Proceedings of the european conference of the phm society 2018* (Vol. 4, p. 403).

Nejjar, I., Geissmann, F., Zhao, M., Taal, C., & Fink, O. (2024). Domain adaptation via alignment of operation profile for remaining useful lifetime prediction. *Reliability Engineering & System Safety*, *242*, 109718.

Qian, Q., Qin, Y., Luo, J., Wang, Y., & Wu, F. (2023). Deep discriminative transfer learning network for cross-machine fault diagnosis. *Mechanical Systems and Signal Processing*, *186*, 109884.

Qian, Q., Wang, Y., Zhang, T., & Qin, Y. (2023). Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis. *Knowledge-Based Systems*, *276*, 110748.

Ramírez-Sanz, J. M., Maestro-Prieto, J.-A., Arnaiz-González, Á., & Bustillo, A. (2023). Semi-supervised learning for industrial fault detection and diagnosis: A systemic review. *ISA transactions*.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., … Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, *109*(5), 756–795.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=uXl3bZLkr3c

Wang, J., Qiu, K., Liu, W., Yu, T., & Zhao, L. (2018). Unsupervised-multiscale-sequential-partitioning and multiple-svdd-model-based process-monitoring method for multiphase batch processes. *Industrial & Engineering Chemistry Research*, *57*(51), 17437–17451.

Wang, Q., Fink, O., Van Gool, L., & Dai, D. (2022). Continual test-time domain adaptation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7201–7211).

Wang, Q., Michau, G., & Fink, O. (2019). Domain adaptive transfer learning for fault diagnosis. In *2019 prognostics and system health management conference (phm-paris)* (pp. 279–285).

Yan, P., Abdulkadir, A., Luley, P.-P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2024). A comprehensive survey of deep transfer learning for

anomaly detection in industrial time series: Methods, applications, and directions. *IEEE Access*.

Zhai, S., Cheng, Y., Lu, W., & Zhang, Z. (2016). Deep structured energy based models for anomaly detection. In *International conference on machine learning* (pp. 1100–1109).

Zhang, J., Zou, J., Su, Z., Tang, J., Kang, Y., Xu, H., ... Fan, S. (2022). A class-aware supervised contrastive learning framework for imbalanced fault diagnosis. *Knowledge-Based Systems*, *252*, 109437.

Zhang, Z., & Deng, X. (2021). Anomaly detection using improved deep svdd model with data structure preservation. *Pattern Recognition Letters*, *148*, 1–6.

Counterfactual Explanation for Auto-Encoder Based Time-Series Anomaly Detection

Abhishek Srinivasan[1, 3], Varun Singapura Ravi[1, 4], Juan Carlos Andresen[1] and Anders Holst [2,3]

[1] *Scania CV AB, Södertälje, Sweden*
abhishek.srinivasan@scania.com

[2] *RISE AB, Stockholm, Sweden*

[3] *KTH Royal Institute of Technology, Stockholm, Sweden*

[4] *Linköping University, Stockholm, Sweden*

ABSTRACT

The complexity of modern electro-mechanical systems require the development of sophisticated diagnostic methods like anomaly detection capable of detecting deviations. Conventional anomaly detection approaches like signal processing and statistical modelling often struggle to effectively handle the intricacies of complex systems, particularly when dealing with multi-variate signals. In contrast, neural network-based anomaly detection methods, especially Auto-Encoders, have emerged as a compelling alternative, demonstrating remarkable performance. However, Auto-Encoders exhibit inherent opaqueness in their decision-making processes, hindering their practical implementation at scale. Addressing this opacity is essential for enhancing the interpretability and trustworthiness of anomaly detection models. In this work, we address this challenge by employing a feature selector to select features and counterfactual explanations to give a context to the model output. We tested this approach on the SKAB benchmark dataset and an industrial time-series dataset. The gradient based counterfactual explanation approach was evaluated via validity, sparsity and distance measures. Our experimental findings illustrate that our proposed counterfactual approach can offer meaningful and valuable insights into the model decision-making process, by explaining fewer signals compared to conventional approaches. These insights enhance the trustworthiness and interpretability of anomaly detection models.

1. INTRODUCTION

Modern electrical and mechanical systems are increasingly equipped with more sensors, enabling the development of new anomaly detection methods to identify and alert on deviations indicating failures or malfunctioning. Traditionally, these anomaly detection systems were meticulously designed for specific machines and specific components. However, this requires deep domain knowledge and understanding of the systems.

Recent data-driven approaches offer a compelling alternative. They leverage generalizable algorithms that can learn from data, eliminating the need for expert-crafted rules for each specific scenario. This reduces the efforts required for building an anomaly detector. Neural networks, in particular, have shown remarkable effectiveness in anomaly detection for various applications (Schmidl, Wenig, & Papenbrock, 2022).

Detecting anomalies in a system using sensor data is a task within the field of multivariate time-series analysis. Current trends of neural-network based time-series anomaly detection methods fall under two main categories, i.e., forecast and reconstruction (Schmidl et al., 2022). The forecasting methods are state-based models, they learn the inherent mechanism for forecasting the future states. When the observations and model forecast deviate by a certain threshold an alarm is raised. On the other hand, the reconstruction-based methods learn to compress the normal data (fault free) to a lower dimensional latent space. This lower dimensional latent space is transformed back to the original space. Any data with the reconstruction error higher than a given threshold is considered anomalous.

In real settings just raising anomaly alert is not enough to act upon it. A context is required, such as to know why the model

is flagging an anomaly and which sensor data is behaving anomalous. Neural networks are inherently black-box models and neural-network-based anomaly detection does not naturally provide its internal decision-making process. Significant progress has been done within the field of explainability in this direction (Molnar, 2020). The explainability methods can provide global or local explanations. The global explanations aim to distill the model in an easily understandable logic form (i.e., to explain the model mechanism). Whereas the local explanations aim to explain the prediction of each input sample, e.g., Saliency map and counterfactuals.

Counterfactual explanation is a promising tool that provides context to the anomalies found by neural-network-based models. This explanation method is especially interesting for diagnostic applications, as their explanation focuses on answering the question: 'why is sample A classified as an anomaly and not normal?'. The usual approach for building counterfactual explanations is to start from an anomalous sample and optimise it via a cost function, towards a counterfactual sample which would be classified as normal by the same model that classified it as anomalous. To our knowledge, there is very limited amount of work focused on explaining time series anomaly detection (Haldar, John, & Saha, 2021; Sulem et al., 2022). From the perspective of component diagnostics and maintenance, the existing approaches have a crucial limitation: they often modify all features within a time series to explain the anomaly. The freedom of adjusting just any signal of the anomalous sample in the optimisation process to change the classification averages out valuable information and spreads it over many signals. This loss of information makes it more difficult to interpret the generated counterfactual and makes it less useful for root-cause analysis and diagnostics.

For gaining valuable insights into the anomalies, it is crucial to know the specific features responsible for the anomaly *and* the reason behind the model's classification. As discussed in the previous paragraph, conventional counterfactual explanations solely address the reason behind the anomalies. In this work, we propose an explanation method that identifies the relevant features *and* simultaneously explains the reason behind the anomaly detection for time series reconstruction-based models.

Our approach was tested on the SKAB benchmark data (Katser & Kozitsin, 2020) and on a real-world industrial time-series data using Auto-Encoder based anomaly detection. The results show that counterfactual explanations, using the proposed approach, provide insightful explanations about the nature of the anomalies such as correlation loss and data drift.

2. Related Work

Counterfactual explanation approaches in general have different focuses, including generating valid, sparse, actionable, and causal explanations (Verma, Dickerson, & Hines, 2020). Few address the problem of explaining time-series or anomaly detection. Haldar et al. (2021) investigate the challenge of generating robust counterfactuals for anomaly detection. They define robust counterfactuals as counterfactual samples that don't flip back to the original class in the vicinity of a certain distance. They solve this by adding a constraint in the cost function used for counterfactual optimisation. Sulem et al. (2022) build upon the previous work DiCE (Mothilal, Sharma, & Tan, 2020) for generating diverse counterfactual bounds for time-series anomaly detection. They promote diversity on the generated counterfactual to address the problems of classical counterfactual explanation methods, i.e., generating only one of many possible solutions. Here, their focus was to provide explanation bounds through diverse explanations.

Other research, such as that by Li, Zhu, and Van Leeuwen (2023) and Antwarg, Miller, Shapira, and Rokach (2021), utilise feature importance, a different class of explanations, for Auto-Encoder based anomaly detection. In contrast to ours, their studies do not target time-series data. Antwarg et al. (2021) use a Shapley-values-based approach (feature importance) for Auto-Encoders to explain the impact of a certain feature on other features reconstruction. (Chakraborttii & Litz, 2020) use feature level thresholds for explanations and use feature selection to raise alarms individually. However, they do not explain the reason behind the model prediction.

To our knowledge, previous work has focused on providing either the relevant features or the reason behind anomaly detection. Whereas our approach provides both; the relevant features responsible for the anomaly and the reason why the model classified it as an anomaly. These two factors play a vital role in planing a meaningful action for diagnostics, such as troubleshooting and maintenance scheduling.

3. Preliminaries

3.1. Auto-Encoder (AE)

Auto-Encoders (AE) are unsupervised modeling approaches. An AE model reduces the input, i.e., high dimensional data $x \in \mathbb{R}^n$ into a low dimensional latent representation (encoding) $z \in \mathbb{R}^k$, where $k < n$, using an encoder $E(x, w_e)$. This encoder is followed by a decoder $D(z, w_d)$ which reconstructs the input (decoding) $\hat{x} \in \mathbb{R}^n$ from the latent representation. The encoder and decoder are neural networks with parameters w_e and w_d, respectively. The training process optimises the parameters of the encoding and decoding functions to provide a reconstruction $\hat{x}$ as close as possible to the input x. Some common loss functions utilised are mean square error (MSE), mean absolute error (MAE), and Huber loss.

To extend the AE approach to time-series data we use convo-

lution-based architectures for the encoder and the decoder. We pre-process the data into time-windows. A time-window of length l is represented as $X = (x_t, ..., x_{t+l}) \in \mathbb{R}^{n \times l}$, where $x_t \in \mathbb{R}^n$ are the signal values at time t.

3.2. Gradient based Counterfactual Explainer

In this section, we outline the fundamental principles of gradient based counterfactual explanation techniques. Counterfactual explanations are generated by gradient optimisation on the objective function posed by Wachter, Mittelstadt, and Russell (2017). The objective function $l(x')$ written in general form is given by

$$l(x') = cost(x', model(x')) + (\lambda * d(x, x')) \,, \quad (1)$$

where x is the sample, x' is the generated counterfactual, λ is the weighted factor and the function $d(.,.)$ is a distance measure. This objective function contains two parts, the first part optimises to flip the class (from anomalous to non-anomalous) of the provided anomalous sample and the second minimizes the change between the explanation and the provided sample. Other custom parts can be added depending on the use-case.

In addition to requiring an objective function, this approach also requires the model to be differentiable to be able to use a gradient-based optimisation for counterfactual generation. A simple gradient descent optimisation is given by

$$x'_i = x'_{i-1} - \eta . \nabla l(x'_{i-1}) \,, \quad (2)$$

where i is the optimisation iteration number, η is the step length and x'_{i-1} is the sample form the previous iteration.

4. METHOD

Our approach has three different modules; illustrated in figure 1: 1) Anomaly detector, 2) Feature selector, and 3) Counterfactual explainer. The anomaly detector detects the anomalies. If the provided sample is anomalous, the feature selector provides a list of relevant features to be explained. The counterfactual explainer builds an explanation on the relevant signals that the feature selector selects.

The anomaly detector (module 1) uses an AE, with an encoder *E* and a decoder *D*. The encoder *E* consists of 1D-convolution layers followed by fully connected layers, whereas the decoder *D* uses a mirrored architecture starting with fully connected layers and then 1-D transpose convolution layers. The resulting outputs from the decoder have the same dimension as the inputs. The AE is trained to minimize the reconstruction loss using Huber loss given by

$$L(Y) = \frac{1}{M} \sum_{ij} \begin{cases} 0.5 \cdot y_{ij}/\beta, & \text{for } \sqrt{y_{ij}} < \beta \\ \sqrt{y_{ij}} - 0.5 \cdot \beta, & \text{otherwise} \end{cases} \,, \quad (3)$$

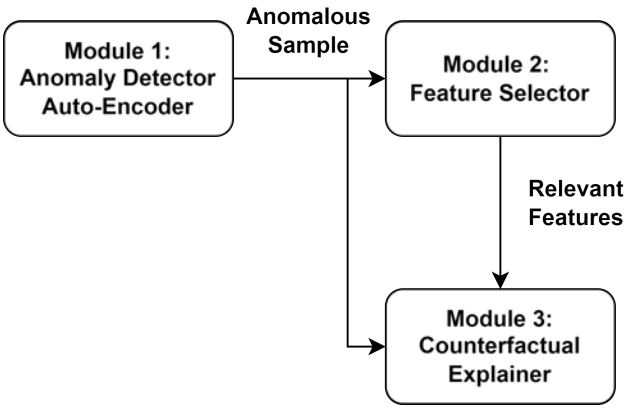


Figure 1. Our proposed methods has 3 modules, 1) Anomaly detector, 2) Feature selector, and 3) Counterfactual explainer. The samples that are classified anomalous by the anomaly detector (module 1) are explained though the feature selector (module 2) and the counterfactual explainer (module 3). The explainer (module 3) uses the selected features from the feature selector and the input sample.

where $Y = (X - \hat{X})^2_\circ \in \mathbb{R}^{n \times l}$, the $\circ$ denotes element wise operation, M is the number of elements of the matrix Y, X is the input to AE and $\hat{X}$ is the reconstruction from AE. Once the AE is trained, the anomaly score (AS) for the validation set is calculated using

$$AS(X, \hat{X}) = MSE(X, \hat{X}) + MAE(X, \hat{X}) \,, \quad (4)$$

where $\left\{X, \hat{X}\right\} \in \mathbb{R}^{n \times l}$ and $MSE(\cdot, \cdot)$ is the mean square error and $MAE(\cdot, \cdot)$ is the mean absolute error of all elements of the matrices. The mean squared error (MSE) element emphasizes larger errors (greater than one) more heavily than the mean absolute error (MAE). Conversely, MAE penalizes smaller errors (below one) more severely. This combination of properties contributes to the effectiveness of the AS. Scores above a threshold are considered anomalous, where the threshold is defined as $\theta_{th} = \mu_{scr} + (k * \sigma_{scr})$ and μ_{scr} is the mean anomaly score on the validation set, σ_{scr} is the standard deviation of anomaly scores on the validation set and k a parameter.

Explanations are provided by the next two modules only when a given sample is classified as anomalous, i.e., when the AS is above the defined θ_{th}. The feature selector (module 2) selects features relevant to the anomaly. It processes the anomalous time window and identifies the features as having either a high or low impact on the anomaly. High-impact features are defined as the ones that are over $m \times percentile(ASW, 90)$ for more than 90% for the window duration, where we choose $m = 0.75$ and ASW is the anomaly score for each feature and time point in the window and is given by

$$ASW(X, \hat{X}) = (X - \hat{X})^2_\circ + |X - \hat{X}|_\circ \,, \quad (5)$$

where $\left\{X, \hat{X}\right\} \in \mathbb{R}^{n \times l}$ and the $\circ$ denotes element wise operation. The key difference between equation (4) and equation (5) lies in the averaging of the error term. ASW in Equation (5) does not average the error, retaining the time and feature dimension assists feature selector to select the right features where anomalies are observed.

The counterfactual explainer (module 3) takes in an anomalous time-window and the features selected by the feature selector. The counterfactual generator uses a modified gradient based explanation (see section 3.2). The difference is that the counterfactual explanation is generated only for the selected features by module 2. This is done by setting the gradients of non-selected features to zero and using the same equation (2) for optimisation, where the $cost$ term is given by the AS in equation (4) and the $model$ given by the anomaly detection AE model.

4.1. Evaluation Metrics

4.1.1. Anomaly Detection Evaluation

As a sanity check, the developed anomaly detection is evaluated with three different metrics; F1-score, False Positive Rate (FPR) and Recall. Equations for these evaluation measures are provided by

$$\text{F1-score} = \frac{TP}{TP + (0.5 * (FP + FN))}\,, \tag{6}$$

$$\text{FPR} = \frac{FP}{FP + TN}\,, \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN}\,, \tag{8}$$

where, TP, FP, TN, and FN refer to true positive, false positive, true negative, and false negative, respectively.

4.1.2. Explainability Evaluation

The developed explainability approach is evaluated with measures: *validity, sparsity, and distance*. *Validity* checks if the generated counterfactual is valid, i.e., if the produced counterfactual is classified as normal. *Sparsity* measures the proportion of features changed in order to generate the counterfactual. Finally, the *distance* provides the mean absolute error distance between the sample and counterfactual.

$$\textit{validity}(x') = \frac{1}{N}\sum_{i=1}^{N} \chi\left(AS(x'_i, AE(x'_i)) < \theta_{th}\right)\,, \tag{9}$$

$$ind(x, x') = \chi(\frac{1}{l}(\sum_{j=1}^{l} |x_{ijk} - x'_{ijk}|) > \epsilon)\,,$$

$$sparsity(x, x') = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{k=1}^{n} ind(x_{ijk}, x'_{ijk})\right)\,, \tag{10}$$

$$d(x, x') = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{l \cdot n}\sum_{j,k} |(x_{ijk} - x'_{ijk})|\right)\,, \tag{11}$$

where $\{x, x'\} \in \mathbb{R}^{N \times n \times l}$

- N: the number of samples,
- l: the sequence length, i.e., the number of time steps per sequence,
- n: the number of features,
- x: sample to be explained,
- x': the counterfactual explanation,
- θ_{th} the threshold used for anomaly detector,
- ϵ: limit defining significant change.
- $\chi(c)$: the indicator function, returning 1 when its argument condition c is true, and 0 otherwise.
- $AE(c)$: is the Auto-Encoder model.

The significant change ϵ in sparsity allows some wiggle room. Typically, this parameter is defined based on the context and the application. In this study ϵ is set to 0.005, i.e., any change above is counted to be a significant.

5. Experimental Setting

5.1. SKAB dataset

(Katser & Kozitsin, 2020) designed a benchmark dataset for time-series anomaly detection. This data is collected from a test-rig consisting of a water tank, valves, and a pump. In this setup, the pump is specifically crafted to extract water from the tank and subsequently circulate it back into the same tank. This setup is equipped with numerous sensors like accelerometer on the pump, pressure sensor after the pump, thermocouple in water, current, and voltage, in total of 8 signals. The collected data is organised in four parts 'no faults', 'valve 1', 'valve 2', and 'others'. 'No fault' has data from normal operation. Data in 'valve 1' and 'valve 2' has data where the corresponding valves were closed for partial duration. The 'others' comprises data from multiple anomaly categories including rotor imbalance, cavitation, and fluid leaks. Each file in 'valve 1', 'valve 2' and 'others' is part normal and part anomalous. It is crucial to note that no two anomaly types co-occur at the same time. The data utilization from different parts of the dataset is summarised in the table 1. The files

$1-4$ are omitted as the data is marked to be simulated and has different characteristics than the other files. After pre-processing into windows, the size of train, validation and test set is 18584, 4658 and 10426 samples. Out of 10426 test samples 3876 are anomalies.

Dataset	Used as	Files
Anomaly-free	80% Train, 20% Valid	All
Valve 1	80% Train, 20% Valid	Only normal behaviour
Valve 2	80% Train, 20% Valid	Only normal behaviour
Others	Test	5-14

Table 1. Table summarizing utilization of SKAB dataset used in our experiments.

5.2. Real-world industrial Data

A commercial, real-world industrial data was collected from a field truck. This data consists of recordings from sensors during normal and anomalous behaviour. Similar to SKAB data, this industrial data encompasses two anomaly types, with no instances of simultaneous occurrences. Two different anomalies were considered: "correlation loss" and "change in relation". A set of 11 relevant sensor signals were utilised for the experiment. The training and validation processes were conducted using two separate dataset containing only normal data (i.e., no-fault data). The test set involved one no-fault scenario and two anomalous runs, where the anomalies were of a different nature. After pre-processing into windows the number of samples in train, validation and test set is 3231, 1074 and 4355 samples. Out of 4355 test samples 1396 were anomalies.

5.3. Model and Explainer Setup

To pre-process the data, we have used min-max normalisation. This involves using the minimum and maximum values from the train-set to normalise the train, validation, and test sets. The time-series sensor signals were pre-processed into smaller chunks using a sliding window technique, with a window length *l* of 64 over *n* signals, *n* being 8 and 11 for SKAB and real-world data respectively.

Experiments on the SKAB dataset employed a random seed of 125. The AE model consists of: i) Encoder with 2 layers of 1D convolution with 64 and 32 filters, kernel size of 5 and stride of 2, followed by a fully connected layer of 8 units; ii) Decoder consists of a mirrored architecture to the above, starting with a dense layer of size 128 followed by 2 layers of 1D transpose convolution with 32 and 8 filters, kernel size of 5 and stride of 2. The model was trained for 150 epochs with a batch size of 64, using Adam optimiser with a learning rate of $\lambda = 0.001$, parameters $\beta 1 = 0.9$, and $\beta 2 = 0.999$, we set $k = 8$ for calculating θ_{th}.

Experiments on the industrial employed uses a random seed of 42. The AE model consists of: i) Encoder with 2 layers of 1D convolution with 32 and 64 filters, padding 1, kernel size of 5 and stride of 1, followed by 4 fully connected layers with 64, 32, 16, and 8 units; ii) Decoder consists of the mirrored architecture, starting with 2 dense layers of size 16 and 32, followed by 2 layers of 1D transpose convolution with 64 and 32 filters, kernel size of 5 and stride of 1. The model was trained for 100 epochs with a batch size of 32, using Adam optimiser (AMSGrad variant) with a learning rate of $\lambda = 0.001$, parameters $\beta 1 = 0.9$, and $\beta 2 = 0.999$, we set $k = 10$ for calculating θ_{th}.

Experiments on both dataset used gradient descent optimisation for 75k iterations, with a learning rate of 0.01 for generating explanations in the the counterfactual explainer (module 3).

6. Results and Discussion

This section is organized into two parts, first evaluation of the anomaly detection and second the results from the counterfactual explanations.

6.1. Results from Anomaly detection

Two AE models were trained, one for each dataset (SKAB and industrial). The anomaly detection threshold was calculated on the validation set, as described in Section 4. The trained models were then evaluated on their respective test sets. The performance of the anomaly detector is summarized in Table 2.

The SKAB dataset results show satisfactory performance with F1-score and Recall around 0.7, along with a False Positive Rate (FPR) of 0.24. The industrial dataset exhibits exceptional performance, achieving F1-score and Recall close to 0.9, with a perfect zero FPR. Anomaly detection confusion matrix for both datasets can be found in Appendix A1.

Table 2. Evaluating anomaly detection models on SKAB and industrial dataset.

Dataset	**F1-score**	**Recall**	**FPR**
SKAB	0.68	0.72	0.24
Industrial data	0.94	0.88	0

6.2. Results from counterfactual Explanation

To demonstrate the effectiveness of our method in explaining time-series anomalies, we compare it with two other approaches:

- **Reconstruction**: This method directly uses the AE reconstruction as the explanation for an anomaly. This is based on the assumption that the reconstructions are projected onto the normal space, hence, a plausible counterfactual explanation.

- **Counterfactual Explainer** (Without Feature Selection): This approach utilizes a counterfactual explainer (module 3) to generate explanations directly for all features, similar to gradient-based counterfactual explanations with $\lambda = 1$ in equation (1). This essentially explains every feature without any selection.
- **Our Proposed Approach** (With Feature Selection): This combines a feature selector (module 2) and a counterfactual explainer (module 3). The feature selector identifies the most relevant features, and the counterfactual explainer then focuses its explanation on these selected features only, with $\lambda = 0$ in equation (1).

We evaluate the explanations generated by these three approaches using three metrics: validity, sparsity, and distance. These metrics are explained in detail in section 4.1.2. The results of this comparison are presented in Table 3.

Table 3. Compilation of evaluation measures from SKAB and industrial dataset. The arrow direction indicates if higher or lower values that makes the approach better.

Dataset	Method	Validity ↑	Sparsity ↓	distance ↓
SKAB	Reconstruction	1.0	1.0	0.246
SKAB	Counterfactual	0.72	1.0	0.214
SKAB	**Ours**	0.67	0.16	0.150
Industrial data	Reconstruction	1.0	1.0	0.140
Industrial data	Counterfactual	0.93	0.99	0.200
Industrial data	**Ours**	0.99	0.17	0.156

Table 3 shows that our approach has reasonably good validity and distance values compared to the other two simpler methods, but with a much better sparsity values than the other methods. Note that the reconstruction method will always have the highest possible validity value due to its nature that the reconstructions are in the same manifold as training data. So this method scores best in this validity measure on both datasets. The counterfactual explainer (without feature selection) has higher validity measure than our proposed method on the SKAB data. The counterfactual explainer (without feature selection) has an advantage of being able to vary all features to provide explanations. This does not necessary mean that the explanation will be more meaningful as by adjusting all features simultaneously the information (the reasons) about the raised anomaly gets diluted. Additionally, altering all signals by the counterfactual explainer (without feature selection) results in a the sparsity scores much worse than our method (with feature selection). Scores form our approach are consistently good in all three measures. To look further into the meaningfulness of the given explanations we illustrate some scenarios in section 6.2.1.

We leverage UMAP embedding (a dimension reduction technique) to achieve two objectives: visualize the relationship between the generated counterfactuals and the test data, and evaluate the *validity* of the explanations independent of the model used for counterfactual generation. In Figure 2 we visualize the UMAP embedding trained on the test-set data from the industrial dataset. Green points represent the non-faulty data (based on ground truth), red points represent the anomalies (based on ground truth), and yellow points represent the projected counterfactuals (generated form our approach). As evident from the Figure 2 the majority of counterfactuals projected on top of the green normal data embeddings, indicating that they represent valid non-faulty behaviors. Only a few, 12 out of 1350 explanation are non-valid (which is reflected in the *validity* measure). These non-valid samples are projected onto the same space as the red faulty data embedding. The lack of valid explanations can be due to parameter selection, optimisation budgets and quality of the feature selection. The validity in confusion-matrix form for the SKAB test data is given in Table 6 in Appendix A2, the validity confusion-matrix form for real-world industrial test data is given in Table 7 in Appendix A2.

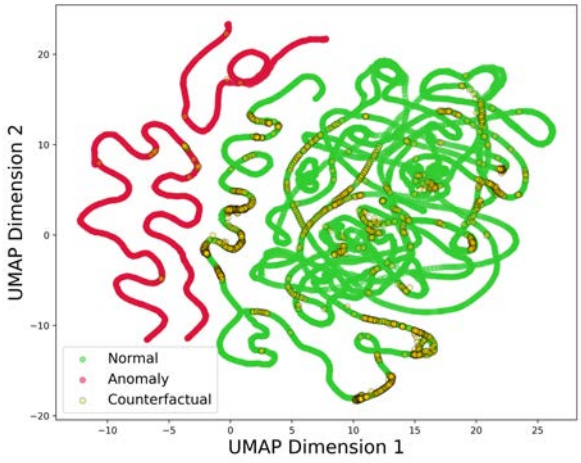


Figure 2. Industrial data: UMAP embedding learnt on no-fault and anomalous data from the test set. Later the generated counterfactual is projected into the same embedding.

6.2.1. Plots showing insights on the explanations

In this section, we show two different explanation scenarios, one from the industrial and the other from the SKAB dataset. Scenario 1 is from the industrial dataset and is illustrated in the Figure 3. The time-window plotted in Figure 3 was classified as anomalous and signal 7 was selected as high impact feature. In Figure 3 we show the input signal 7 and signal 8 in blue and the counterfactual explanation in orange (see Figure 6 in the Appendix A3 for comparison with reconstruction and counterfactual signals). The root cause of this anomaly is a loss of correlation in signal 7. In normal (no-fault) data signal 7 and signal 8 are correlated with a median correlation coefficient of 0.99 and our explanation restored the correlation between the signals on the anomalous data (of this type)

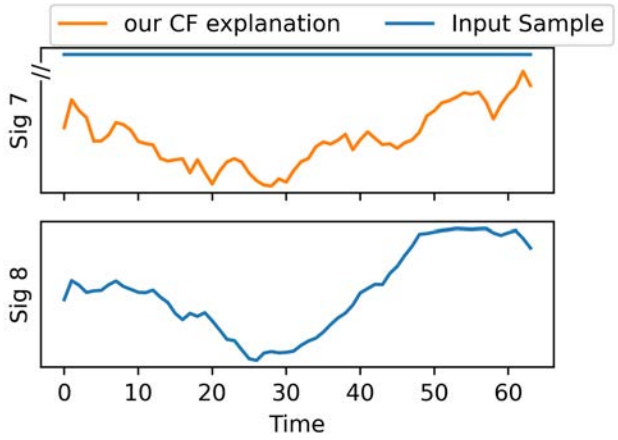


Figure 3. Plot of counterfactual explanation generated by our approach for industrial dataset. This plotted sample was of correlation loss anomaly. Signal 7 and signal 8 in blue show the input and signal 7 in orange shows the explanation.

to a median correlation coefficient of 0.93.

Figure 4 shows the second scenario from SKAB data. Here the selected anomalous window belongs to the rotor imbalance anomaly. This window was classified as anomaly and our feature selector selected Acc1RMS and Acc2RMS signals which belong to the accelerometer sensors as high impact features. The explanation from our approach indicates that the vibrations observed by the accelerometer should be lower to be classified as normal (see the Figure 5 in Appendix A3 to see the comparison with CF and reconstruction signals).

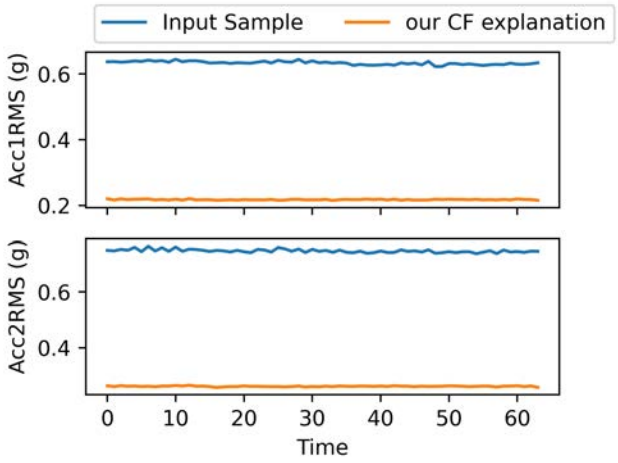


Figure 4. Plot showing the counterfactual explanations provided by our approach and the anomalous samples. Only the high impact features that were explained are plotted.

In Scenario 1, the explanation hints that the the correlation between signal 7 and signal 8 is broken by the flat line and is confirmed by the correlation analysis. Combining this explanation with the domain expertise, it is easy to conclude that the sensor for signal 7 is broken. In Scenario 2 from the explanation we know that we have too high vibrations that often are originated by rotor imbalance. The explanations provided by our approach are meaningful in the context of system functionality and provides insights about the nature of the anomaly when compared to other approaches. This is due to it's capacity to select features for explanation. The comparison between different approaches can be seen in the detail in the Figure 5 and Figure 6 provided in Appendix A3.

7. Conclusion

In summary, our work proposes a method for explaining AE-based anomaly detection for time-series data, based on relevant feature selection and counterfactual explanations. This approach can answer on which features the anomaly is located together with why the sample was classified as an anomaly. We find that these explanations have consistently good scores in all three measures, *validity*, *sparsity* and *distance*, which translates into useful and actionable insights from a diagnostic perspective. We give two examples, one from a benchmark dataset and one from an industrial dataset, on how the proposed method can help to diagnose the classified anomalies from the AE anomaly detection model. This contribution serves as a diagnostic tool, enhancing our understanding and analysis of anomalous events. Note that the quality of explanation depends on the performance of the selected anomaly detection model, parameter selection and the quality of feature selection.

Future work can focus on different optimisations for the explanation, improve the quality of the feature selector and understand the model relation with the explainer.

Acknowledgment

The project was funded FFI Vinnova under project number 2020-05138. We thank Deepthy and Swathy for good discussions.

References

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert systems with applications*, *186*, 115736.

Chakraborttii, C., & Litz, H. (2020). Improving the accuracy, adaptability, and interpretability of ssd failure prediction models. In *Proceedings of the 11th acm symposium on cloud computing* (pp. 120–133).

Haldar, S., John, P. G., & Saha, D. (2021). Reliable counterfactual explanations for autoencoder based anomalies. In *Proceedings of the 3rd acm india joint international conference on data science & management of data (8th*

acm ikdd cods & 26th comad) (pp. 83–91).

Katser, I. D., & Kozitsin, V. O. (2020). *Skoltech anomaly benchmark (skab).* `https://www.kaggle.com/dsv/1693952`. Kaggle. doi: 10.34740/KAGGLE/DSV/1693952

Li, Z., Zhu, Y., & Van Leeuwen, M. (2023). A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *18*(1), 1–54.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, *15*(9), 1779–1797.

Sulem, D., Donini, M., Zafar, M. B., Aubet, F.-X., Gasthaus, J., Januschowski, T., ... Archambeau, C. (2022). Diverse counterfactual explanations for anomaly detection in time series. *arXiv preprint arXiv:2203.11103*.

Verma, S., Dickerson, J. P., & Hines, K. E. (2020). Counterfactual explanations for machine learning: A review. *ArXiv*, *abs/2010.10596*.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, *31*, 841.

APPENDIX

A1. Confusion Matrix for the Anomaly detector

In this section, the confusion matrices for the anomaly detector on SKAB and real-world industrial dataset are presented in Table 4 and Table 5, respectively.

Table 4. Confusion Matrix for SKAB test data.

		Prediction outcome		
		P	**N**	**Total**
Actual value	**P′**	2788	1088	3876
	N′	1573	4977	6550
	Total	4361	6065	10426

Table 5. Confusion Matrix for real-world industrial test data.

		Prediction outcome		
		P	**N**	**Total**
Actual value	**P′**	1350	171	1521
	N′	0	2834	2834
	Total	1350	3005	4355

A2. Confusion Matrix like expression for validity using our approach

In this section, we show valid samples in a confusion-matrix like setting for SKAB and real-world industrial dataset are presented in Table 6 and Table 7 respectively.

Table 6. Validity confusion Matrix for SKAB test data.

		Prediction outcome		
		Valid	**Not Valid**	**Total**
Model Prediction	**True Positives**	1885	903	2788
	False Positives	1068	505	1573
	Total	2953	1048	4361

Table 7. Validity confusion Matrix for real-world industrial test data.

		Prediction outcome		
		Valid	**Not Valid**	**Total**
Model Prediction	**True Positives**	1338	12	1350
	False Positives	0	0	0
	Total	1338	12	1350

A3. Plot comparing different approaches

A sample from rotor-imbalance anomaly is plotted along with different explanations in the Figure 6. The plotted sample is the same as in the Figure 4. In figure 6, explanations from different methods are compared. It can be seen that other approaches explains by changing all the features where as the explanation from our approach changes only *ACC1RMS* and *ACC2RMS* signals. In similar way , for the sample plotted in the Figure 3, in Figure 5, we compare our approach with other type of explanations.

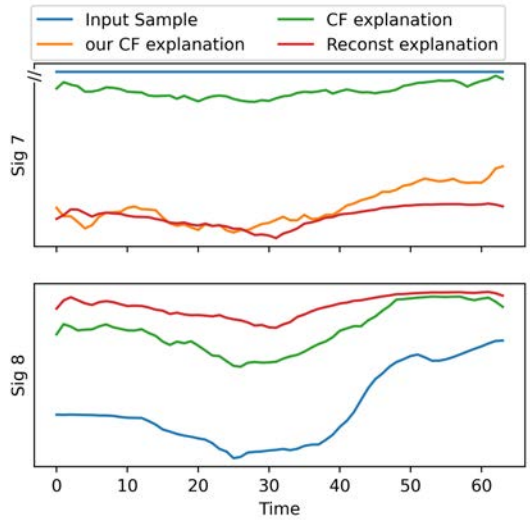


Figure 5. Plot showing the explanations provided by reconstruction, counterfactual(CF) based (i.e., without feature selector) and our approach (i.e., with feature selector). Additionally the input sample is plotted.

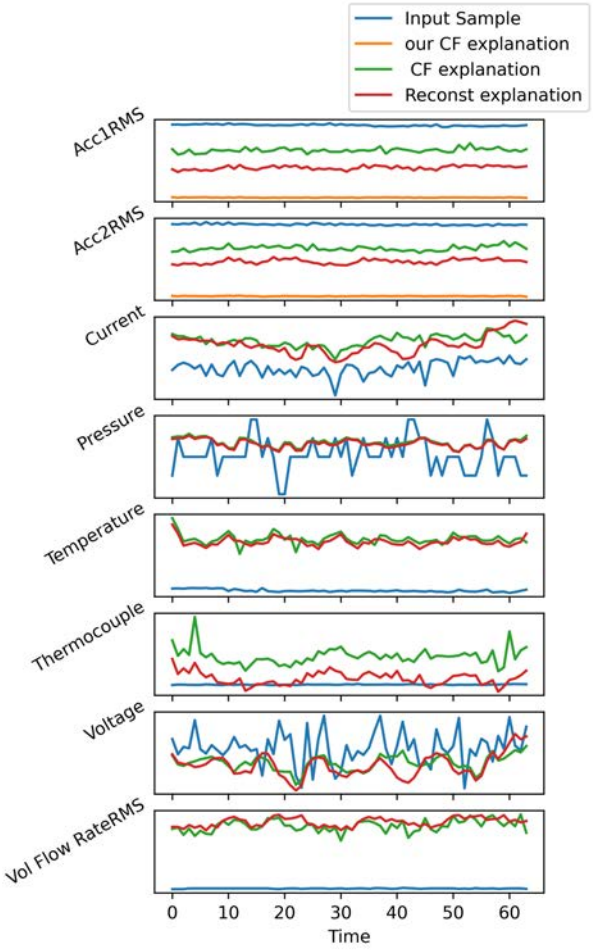


Figure 6. Plot showing the explanations provided by reconstruction, counterfactual(CF) based (i.e., without feature selector) and our approach (i.e., with feature selector). Additionally the input sample is plotted.

Damage Detection using Machine Learning for PHM in Gearbox Applications

Lisa Binanzer, Tobias Schmid, Lukas Merkle and Martin Dazer

Institute of Machine Components, University of Stuttgart, 70569 Stuttgart, Germany

lisa.binanzer@ima.uni-stuttgart.de
st157630@stud.uni-stuttgart.de
lukas.merkle@ima.uni-stuttgart.de
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

Early damage detection in gearbox applications enables the implementation of Prognostics and Health Management (PHM). On the one hand, the earliest possible damage detection provides a precise in-sight into the state of health of a gearbox. In addition, early damage detection offers the possibility to slow down the damage progress and extend the remaining useful life (RUL) by intervening in the operating state at an early damage stage. The main contribution of this work is that existing Machine Learning tools are applied to the challenge of very early damage detection in gearboxes. Thus, the need for complex physically based data evaluation is avoided. The aim of this investigation is a comparison of two different machine learning approaches. To investigate the detection possibilities test bench experiments were conducted with a single stage spur gearbox. For a comprehensive investigation, i.e. to detect damage under different operating conditions, the test runs are carried out at different damage sizes, speeds and torques. Based on the recorded vibration data, the damage detection is examined. Two machine learning approaches of anomaly detection are considered: An encoding approach and a loss approach. The same sparse autoencoder is developed for both approaches Both machine learning approaches are able to detect even the smallest damage of about 0.5 % in most operating states. The loss approach allows the different damage stages to be recognized much more clearly than the encoding approach. The comparison of the different approaches provides valuable insights for the further development of robust damage detection algorithms.

1. INTRODUCTION

In many mobile and stationary applications, gearboxes are essential for adjusting speed and torque. The greater the power that needs to be transmitted, the larger and more expensive the corresponding gear units are. In gearboxes one of the most common types of damage on a tooth flank is pitting. As soon as a pitting exceeds a size of 4 % in relation to the size of the tooth flank, the gear is considered as failed according to the 2016 International Organization for Standardization [ISO] report. Damaged tooth flanks are one of the leading reasons of downtime and each failure can be associated with high repair costs and time-consuming repair work. This particularly applies to large gearboxes and applications in remote locations, such as offshore wind power drives. For this reason, gearboxes in critical industrial applications are often equipped with condition monitoring systems (CMS) based on vibration sensors. They continuously monitor the current state of health of the gearbox. If the CMS detects damage, depending on the damage extent, a complete shutdown or a load reduction can be initiated. Expensive subsequent damage can be prevented and, in case of a reduced load, the remaining useful life (RUL) of the gearbox can be extended until it is repaired or replaced. However, the CMS's are only developing their full potential if damage can be detected at a very early stage.

The earliest possible damage detection in gearboxes enables comprehensive Prognostics and Health Management (PHM) to be implemented in gearbox applications. According to Goebel, Celaya, Sankararaman, Roychoudhury, Daigle and Abhinav (2017), a PHM approach consists of the 5 sub-areas of the system: data, diagnosis, prognosis, optimization and the system itself. The earliest possible damage detection affects all of these areas.

First of all, the health of a system, which according to the 2017 Institute of Electrical and Electronics Engineers [IEEE] committee standards include all information regarding the functionality of a system, can be diagnosed much more precisely using the data of the system. The time gained by early damage detection can be used to acquire more data for a possible RUL prediction. Finally, health management of the system can be realized. Health management describes the control of damage according to the aim of the PHM solution

(Bertsche & Dazer, 2023). A potential goal is the optimal utilization of the RUL without unexpected failure until scheduled maintenance. For instance, this can be achieved by avoiding particularly damaging operating points. Another option is to implement an adaptive operating strategy that enables the extension of the RUL without any loss of performance (Gretzinger, Lucan, Stoll & Bertsche, 2020). Due to the application of the operating strategy, the load on the predamaged tooth is significantly reduced. This results in a slowdown of the damage progress. The other teeth on the circumference of the gear, which can still withstand the designated load, are being slightly overloaded. Thus, the load reduction is compensated without any overall power reduction. Control of the plant is carried out by a corresponding optimization algorithm.

Overall, comprehensive PHM in gearbox applications offers numerous benefits. However, damage detection at a very early stage is a prerequisite. The aim of this study is to investigate the earliest possible damage detection in gearboxes with the help of machine learning. Experiments were conducted on a test gearbox, vibration data was recorded and evaluated using an Autoencoder (AE).

2. Experiments and Machine Learning Approaches

Following, the test bench experiments for the earliest possible damage detection are described first. Subsequently, the developed autoencoder is presented. Finally, the two machine learning approaches (encoding and loss), which are based on the developed autoencoder, are discussed in more detail.

2.1. Test Bench Experiments

The test gearbox is designed as a single stage spur gear unit. Figure 1 illustrates the design of the test gearbox. The gear ratio is $i = 25/36 = 0.69$. More information on the test gears in (Binanzer, Merkle, Dazer & Nicola, 2023).

The test bench is set up as an inline concept (2 motor concept). The electric drive motor loads the transmission input side and the second electric motor loads the transmission output side. Torque measuring shafts and incremental encoders for speed and angular position measurement are mounted between the electric motors and the test gearbox. Further information on the test bench setup can be found in (Binanzer et al., 2023). The mounted test gearbox on the test bench is shown in figure 2.

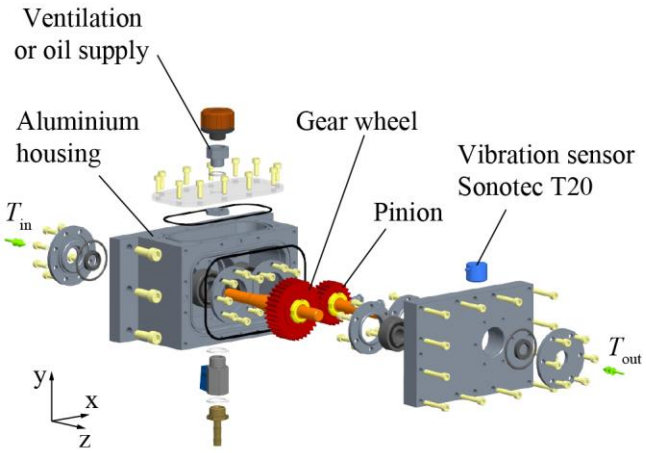


Figure 1. Design of the test gearbox.

For lubricating the tooth contact, FVA reference oil no. 3 (see the 1985 Research Association for Drive Technology [FVA] report) is used. This is a mineral oil without additives with a viscosity corresponding to ISO 3448 (see the report (ISO, 2010)). To ensure constant test conditions, the oil is preconditioned in an external fluid tempering device to 29.54 °C. A gear pump supplies the oil to the tooth contact. A Pt100 temperature sensor in the oil supply measures the oil temperature during the test runs.

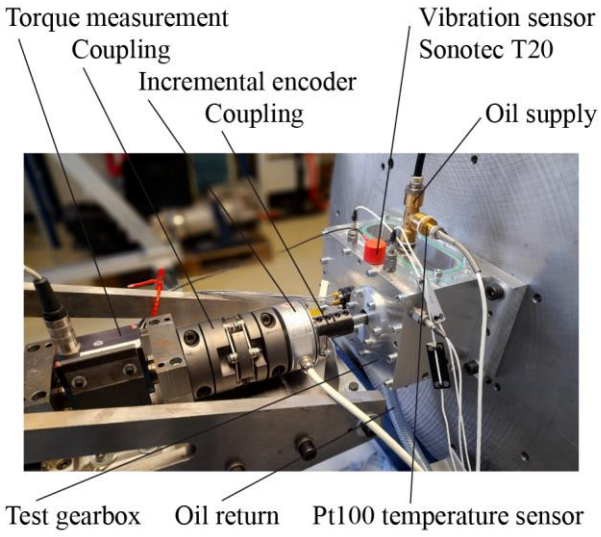


Figure 2. Test gearbox mounted.

The test gearbox is equipped with a Sonotec T20 sensor. This ultrasonic accelerometer is located between the bearings and measures in the y-direction. The maximum measuring frequency of the Sonotec T20 sensor is 100 kHz. However, the sampling rate of the Sonotec T20 sensor is limited to 96 kHz due to the maximum sampling rate of the data acquisition system that is used (PAK MK2). Thus, according to the Nyquist-

Shannon sampling theorem, maximum frequencies of 48 kHz can be measured with the system.

Since the earliest possible damage detection using machine learning algorithms is to be investigated as part of this work, damage well below the 4 % criterion is examined. Artificially generated pitting serves as representative gear damage. Following a test series without damage, a total of three damage sizes are tested - small (S), medium (M) and large (L). The tests without damage serve as a reference and thus as a training data set for the machine learning algorithms. The pitting damage is applied using a numerically controlled milling machine. This ensures that the pitting can be easily and reproducibly manufactured on the tooth flanks. A suitable radius milling cutter with a head diameter of 2 mm is used. Due to the higher number of load cycles, pitting damage usually occurs on the pinion. Consequently, the artificial damage is applied on the pinion. The gear wheel remains undamaged. The pitting is located in the center of a tooth flank below the pitch circle. Figure 3 shows the pinion with the manufactured pitting damage.

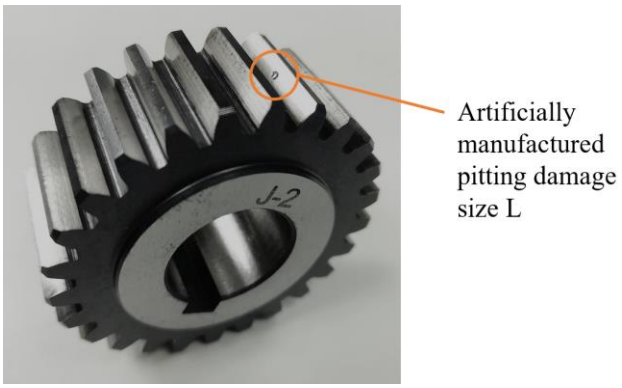


Figure 3. Pinion of second gear pair with manufactured pitting damage size L (1.72 %).

For a comprehensive investigation, i.e. to detect damage under different operating conditions, the test runs are carried out not only at different damage sizes, but also at different speeds and torques. In each of the four test series (no damage, S, M, L), six operating conditions are tested. These six operating states result from a combination of two speed levels (72 rpm, 636 rpm) and three torque levels (18 Nm, 24 Nm, 30 Nm). Within a test series, the six operating states are varied randomly in their sequence. The measurement duration for each test run is 100 s. Between the test series, the damage on the pinion is then artificially applied and milled larger. In this way, increasing damage on the pinion can be tested and the different damage sizes can be directly compared with each other. In order to achieve a more meaningful result, the tests are carried out with a total of three pairs of gears. The pinion and gear are always tested in the same pairs. This ensures that the algorithms do not detect any anomalies caused by manufacturing tolerances or material deviations of different pinion - gear combinations. The only difference between the test series of a gear pair is the increasing damage on the pinion. The surface area of each pitting is measured after the milling process using a digital microscope (see table 1).

Table 1. Pitting surface areas.

Gear pair	Pitting level	Pitting surface in mm^2	Relative surface area in %
1	S	0.48	0.61
1	M	0.97	1.23
1	L	1.40	1.77
2	S	0.40	0.51
2	M	0.92	1.16
2	L	1.36	1.72
3	S	0.62	0.78
3	M	1.01	1.28
3	L	1.40	1.77

2.2. Autoencoder

A test run duration of 100 s and a sample rate of 96 kHz result in 9,600,000 data points per test (acceleration over time). These data points are then converted into the frequency spectrum. For this purpose, Fast Fourier Transforms (FFT) of 10,000 data points each are performed, thus 960 FFT's per test. Each FFT results in 5,001 frequency points. The two rotational speeds of the tests result in the following: At a speed of 72 rpm, approximately 11.6 FFTs are conducted for each revolution of the pinion. At a speed of 636 rpm, approximately 1.3 FFTs are generated for each revolution of the pinion.

The AE developed in this study emerged from literature research and empirical hyperparameter tuning. It is a multilayer AE, which consists of three joined AEs, each with a hidden layer, see figure 4. The number of units in layer l is defined as s_l. The input of the first AE used in this study contains $s_1 = 5{,}001$ units corresponding to the number of frequency points per FFT. Accordingly, layer 3 and 5 have $s_3 = s_5 = 5{,}001$ units. The first hidden layer (layer 2) has $s_2 = 560$ units, the second hidden layer (layer 4) has $s_4 = 200$ units. Finally, the third hidden layer (layer 6) learns a compressed representation of the frequency spectrum with $s_6 = 50$ features. The output after the last decoding process again has $s_7 = 5{,}001$ frequency points. The aim of the AE is to ensure that the output $\hat{x}$ is the most accurate possible reproduction of the input x. For this purpose, the AE has to learn a function $h_{W,b}(x)$ with the parameters W and b, for which the following is valid:

$$h_{W,b}(x) \approx x \qquad (1)$$

Since the AE in this paper has a total of 7 layers, layer 1 corresponds to the input and layer 7 to the output:

$$x = x^{(1)} \tag{2}$$

$$\hat{x} = h_{W,b}(x) = x^{(7)} \tag{3}$$

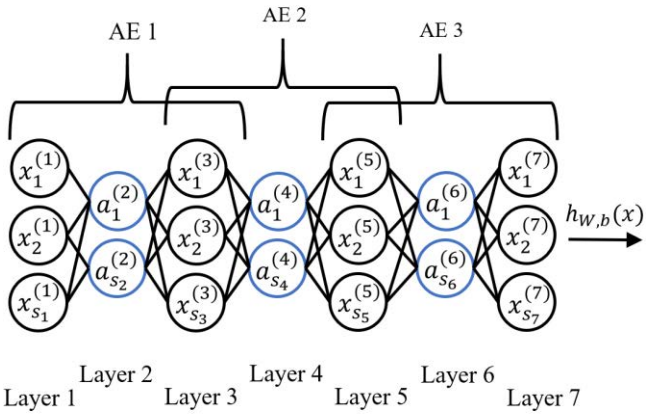


Figure 4. Structure of the Autoencoder.

For parameter W, the notation $W_{i,j}^{(l)}$ is used and this is associated with the weighting of the connection between unit j in layer l and unit i in layer $l+1$ (Ng, 2011). Parameter $b_i^{(l)}$ is the bias associated with unit i in layer $l+1$ (Ng, 2011). Eq. (4) and (5) are valid to the layers of the AE. For this, $a_{s_l}^{(l)}$ corresponds to the output of the hidden unit s_l in layer $l = 2, 4, 6$ and $x_{s_l}^{(l)}$ corresponds to the output of unit s_l in layer $l = 3, 5, 7$.

$$a_{s_l}^{(l)} = f\left(\sum_{i=1}^{s_l-1} W_{s_l,i}^{(l-1)} x_i^{(l-1)} + b_{s_l}^{(l-1)}\right) \tag{4}$$

$$x_{s_l}^{(l)} = f\left(\sum_{i=1}^{s_l-1} W_{s_l,i}^{(l-1)} a_i^{(l-1)} + b_{s_l}^{(l-1)}\right) \tag{5}$$

The function $f(z)$ with $f: \mathbb{R} \to \mathbb{R}$ is called activation function. The sigmoid function (Eq. (6)) is chosen as the activation function in this study. This function can assume values between 0 and 1, see figure 5.

$$f(z) = \frac{1}{1 + exp(-z)} \tag{6}$$

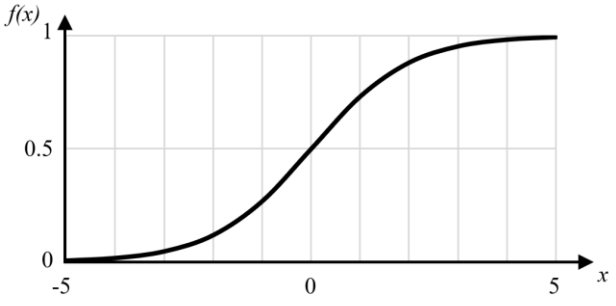


Figure 5. Sigmoid function.

The AE requires training with a training data set with m training examples. The data from the tests without damage serves as the training data set. Thus, the training data set consists of 960 FFTs each ($m = 960$). If training data set number m is used as input, then $x^{(l),(m)}$ results in layer l. The three AEs are trained one after the other. The corresponding loss function $J_{AE}(W, b)$ with $AE = 1, 2, 3$ is defined for training the AE for both approaches (encoding and loss) and consists of three terms:

$$\begin{aligned} &J_{AE}(W, b) \\ &= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\left\|x^{(l+2),(i)} - x^{(l),(i)}\right\|^2\right)\right] \\ &+ \lambda \sum_{l}^{l+1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \\ &+ \beta \sum_{j=1}^{s_l} KL(\rho \| \hat{\rho}_j) \end{aligned} \tag{7}$$

$$with$$

$$l = 1\ for\ AE = 1;$$

$$l = 3\ for\ AE = 2;$$

$$l = 5\ for\ AE = 3$$

The first term of the loss function is the average sum-of-squares error term (Ng, 2011). It describes the deviation of $\hat{x}$ from x and can therefore also be described as a reconstruction error. The second term is the regularization term, also known as the weight decay term, which tends to reduce the size of the weights W and helps to prevent overfitting (Ng, 2011). It therefore ensures that the network does not simply memorize the data, but learns the underlying structure. The weight decay parameter λ determines the relative importance of the second term. $\lambda = 0.001$ is selected.

Additionally, a sparsity constraint is imposed on the hidden units of layer 2, 4 and 6. Therefore a sparse AE is obtained and the third term of the loss function is the sparsity penalty term. With $\beta = 1$, the weighting of the term corresponds to a single weighting compared to the first term. A sparsity constraint permits the AE to learn the underlying structure in the data even with a large number of hidden units. For this

purpose, the average activation of unit j in the hidden layer l is calculated:

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(l)}\left(x^{(i)}\right)\right] \quad (8)$$

This average activation of the neuron should match the selected sparsity constraint ρ:

$$\hat{\rho}_j = \rho = 0.3 \quad (9)$$

The sparsity penalty term penalized if $\hat{\rho}_j$ deviates significantly from ρ. The penalty term is based on the Kullback-Leibler (KL) divergence:

$$\sum_{j=1}^{s_l} KL\left(\rho \| \hat{\rho}_j\right) = \sum_{j=1}^{s_l} \rho \, log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\, log\frac{1-\rho}{1-\hat{\rho}_j} \quad (10)$$

The aim of training the AE is to minimize the function $J(W,b)$ as much as possible. $J(W,b)$ becomes as small as possible when an optimal combination of the parameters W and b is found. First, these parameters are randomly initialized, then backpropagation is applied. During backpropagation, the connections of the AE's units are either strengthened or weakened via the weightings to further minimize the loss between input and output. Each AE is trained for 100 iterations.

2.3. First evaluation approach: Encoding

After training the AE on the basis of the vibration data of a test without damage, the trained AE is used to evaluate all data of an individual operating condition of this gear pair. For this purpose, the recorded vibration data of the test without damage and the three tests with damage sizes S, M and L are appended to each other and an FFT is generated from 10,000 data points each. This results in a total of 3,840 FFTs. These are each encoded to 50 features using the trained AE. A Principal Component Analysis (PCA) is then used to determine a one-dimensional real number from these 50 features. According to the number of FFTs, 3,840 real numbers are obtained. The first 960 one-dimensional representations of the frequency spectrum are those of the test without damage. The data points are normalized between 1 and 2.

The data points of the test without damage are transformed using a Box Cox transformation. The aim is to modify the data in such a way that it is closer to a normal distribution. This provides a standardized baseline for the comparison with the untransformed data from the tests with damage. For the comparison, the arithmetic mean μ and the standard deviation σ are determined from the transformed data without damage. Three intervals are defined based on the standard deviation:

$$1^{st}\text{ interval: } \mu \pm \sigma \quad (11)$$

$$2^{nd}\text{ interval: } \mu \pm 2\sigma \quad (12)$$

$$3^{rd}\text{ interval: } \mu \pm 3\sigma \quad (13)$$

In case of an optimal normal distribution, 68.27 % of the data are in the first interval, 95.45 % in the second interval and 99.73 % in the third interval, see figure 6.

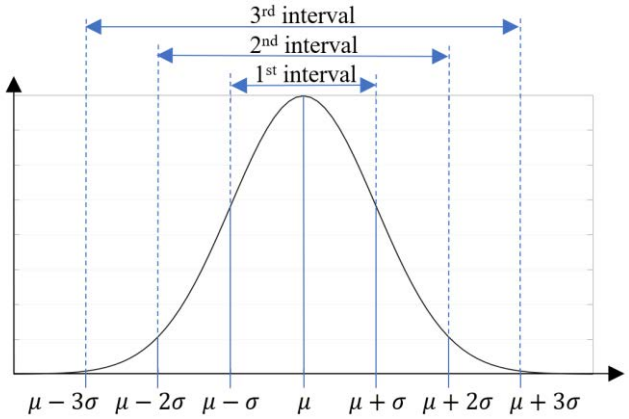


Figure 6. Normal distribution.

Based on the determined intervals, it is calculated how many of the 960 data points each of the tests with damage size S, M and L are outside the intervals. Since the limits of the intervals are further apart as the interval increases, the proportion of data points of the tests with damage that are outside the limits decreases. The first interval will therefore always show a greater deviation between tests without and with damage than intervals 2 and 3.

However, the different machine learning approaches (encoding and loss) can be compared with each other on the basis of the interval method, as it is used in both approaches. If a deviation between the test without and with damage can still be detected with the second or third interval in one approach, the damage is more clearly detectable and the approach is therefore more suitable. The approaches can therefore also be compared in terms of how much the detectability decreases with increasing interval. The less influence the used interval of an approach has on the detectability, the more suitable the approach is.

In the context of this paper, no fixed threshold for the data proportion outside the limits is to be defined with which damage can be detected. Instead, the two approaches (encoding and loss) are to be evaluated and compared with each other based on the predefined intervals.

2.4. Second evaluation approach: Loss

In the second evaluation approach, the AE is also trained on the basis of the vibration data from a test without damage. All vibration data from the test without damage and the three damage variables S, M and L of the individual operating state

are then appended to each other. After 3840 FFTs have been generated from 10,000 data points each, these are encoded and decoded using the trained AE. The FFTs produced by the AE are then compared to the original FFTs by calculating the mean squared error (first term of the loss function, see Eq. 7). The loss approach, such as the encoding approach, provides 960 real numbers per test without damage, damage size S, M and L. Based on this, a Box Cox transformation can be performed again on the data without damage and the three intervals can be determined (see Eq. (11), (12) and (13)). Subsequently, the proportion of data points outside the intervals is calculated for each of the tests with damage.

3. RESULTS

The results of the encoding and the loss approach are presented below.

3.1. Encoding approach

Figures 7 to 12 show the results of the encoding approach of the first gear pair for each operating condition. The lower and upper limits of the first interval are determined on the data without damage (see Eq. (11)). The upper and lower limits are marked with a red line. All blue data points are within this interval, all orange ones outside. For the damage sizes S, M and L, the proportion of data points inside and outside the interval is calculated.

For all operating conditions, it can be seen that the plotted data points per damage size scatter significantly. Overall, this applies even more to the higher speed level of 636 rpm than to the lower speed level of 72 rpm. The most difficult pitting to detect using the encoding approach for the first gear pair is pitting M at 72 rpm and 24 Nm (see figure 8). Here, only 20.2 % of the data points lie outside the first interval. The difference is therefore not as significant as with the other pitting sizes or operating conditions, for which at least 33.5 % of the data points always lie outside the interval.

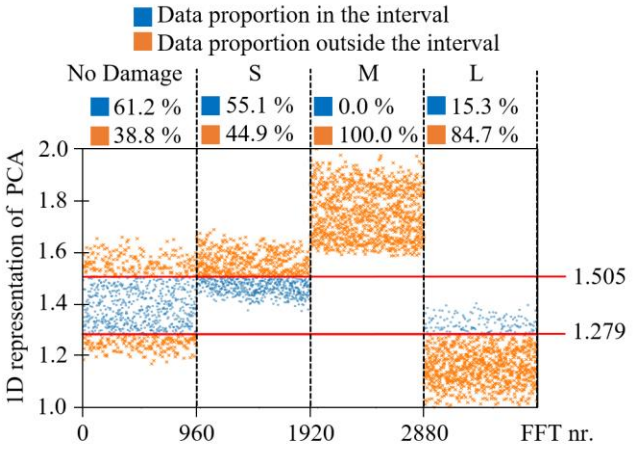


Figure 7. Encoding approach, 1st gear pair, 72 rpm, 18 Nm, 1st interval.

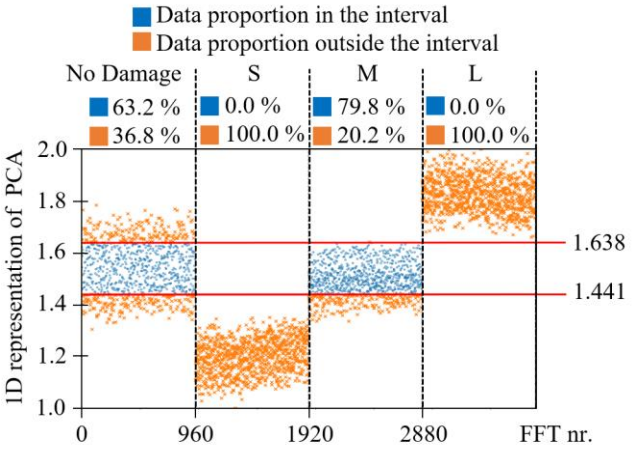


Figure 8. Encoding approach, 1st gear pair, 72 rpm, 24 Nm, 1st interval.

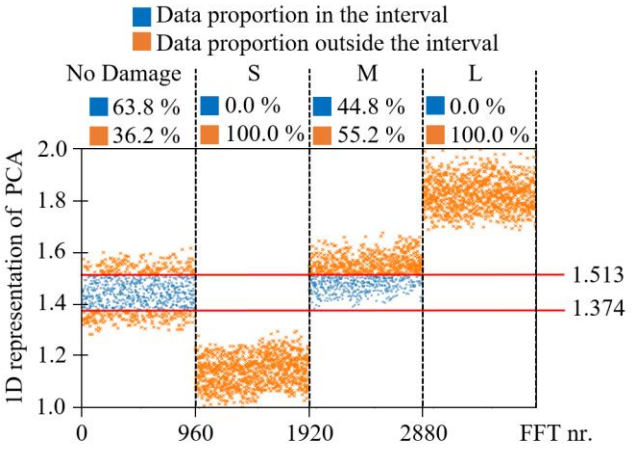


Figure 9. Encoding approach, 1st gear pair, 72 rpm, 30 Nm, 1st interval.

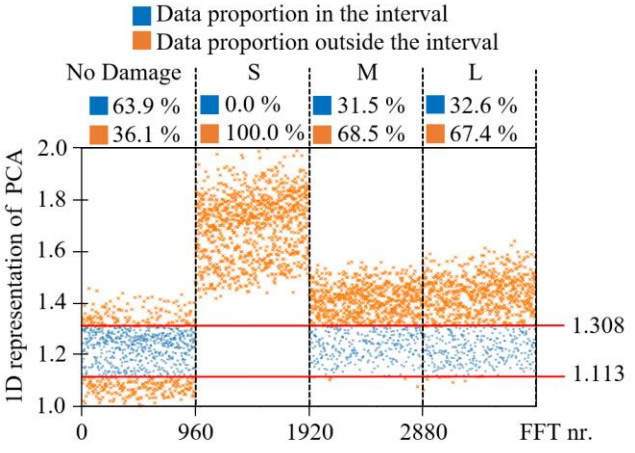


Figure 10. Encoding approach, 1st gear pair, 636 rpm, 18 Nm, 1st interval.

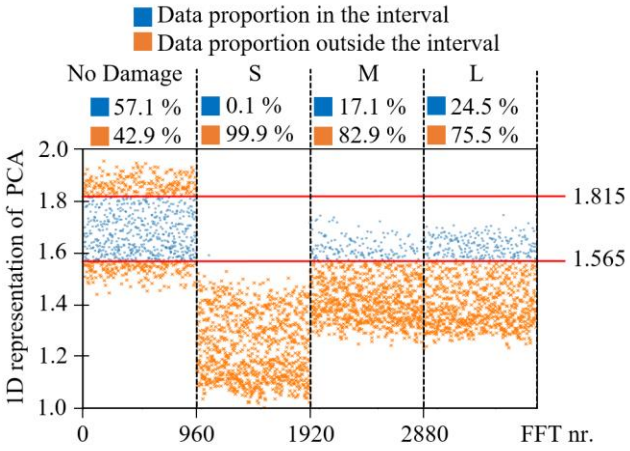


Figure 11. Encoding approach, 1st gear pair, 636 rpm, 24 Nm, 1st interval.

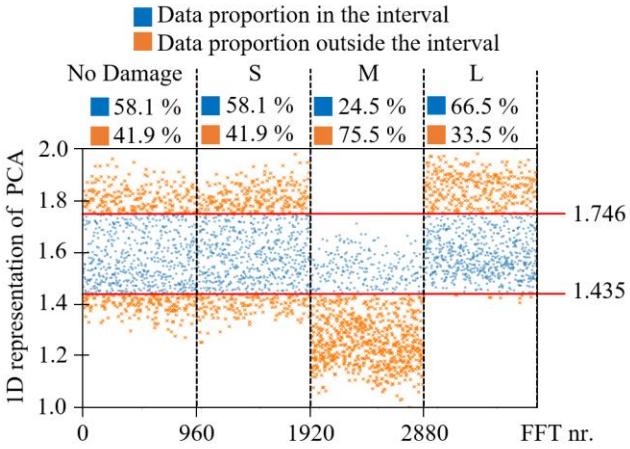


Figure 12. Encoding approach, 1st gear pair, 636 rpm, 30 Nm, 1st interval.

In addition to the encoding evaluation of the first gear pair using the first interval, the second and third intervals are also determined (see Eq. (12) and (13)). As mentioned, this is only used for comparison with the loss approach, as the detectability decreases as the interval increases. However, the aim is to evaluate how much the detectability decreases. For the operating condition 72 rpm and 18 Nm, this results in figure 13 (second interval) and figure 14 (third interval).

While for the first interval only 61.2 % of the data points without damage are within the limits (see figure 7), in the second interval 99.1 % (see figure 13) and in the third interval all data points (see figure 14) are within the limits. As the limits therefore have a greater distance, it is more difficult to detect a difference to the data points of the experiments with damage. Even in the evaluation with the second interval, only 1.7 % of the data points for pitting size S are outside the interval (see figure 13). If the third interval is used for the evaluation, no difference is recognizable, as all data points of pitting S are within the interval (see figure 14). Pitting sizes M and L are also more difficult to detect as the interval increases.

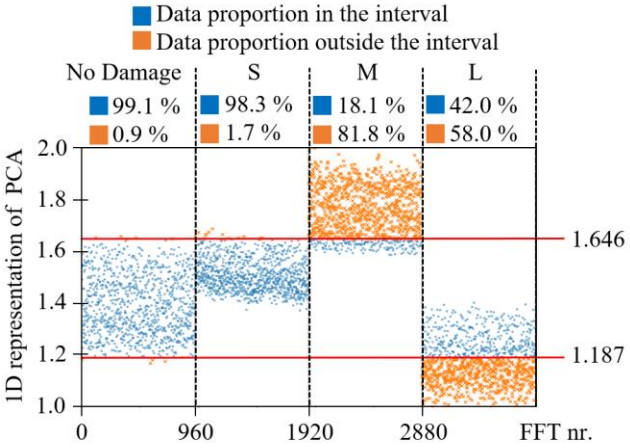


Figure 13. Encoding approach, 1st gear pair, 72 rpm, 18 Nm, 2nd interval.

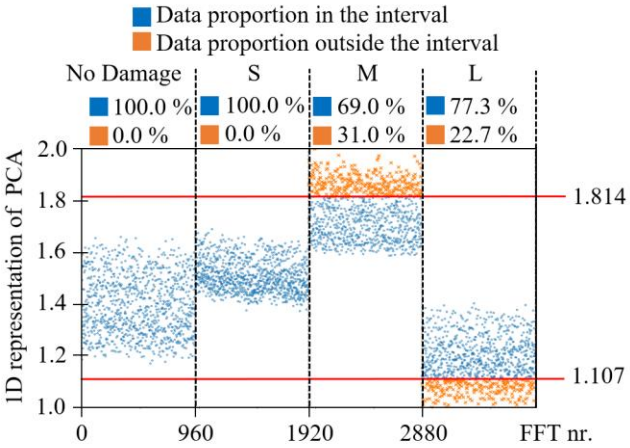


Figure 14. Encoding approach, 1st gear pair, 72 rpm, 18 Nm, 3rd interval.

Figure 15 presents the evaluation of the encoding approach for all operating conditions of the first gear pair. The proportion of data points for damage sizes S, M and L that are outside the respective interval is shown.

When using the second interval, some pitting can no longer be detected (pitting S at 72 rpm and 18 Nm, pitting M at 72 rpm and 24 Nm, pitting S and L at 636 rpm and 30 Nm) or less clearly (13.4 % for pitting M at 72 rpm and 30 Nm). For all other pitting and operating conditions, at least 40.8 % of the data points are always outside the limits of the second interval. If the third interval is used for detection, the detectability of the pitting decreases significantly.

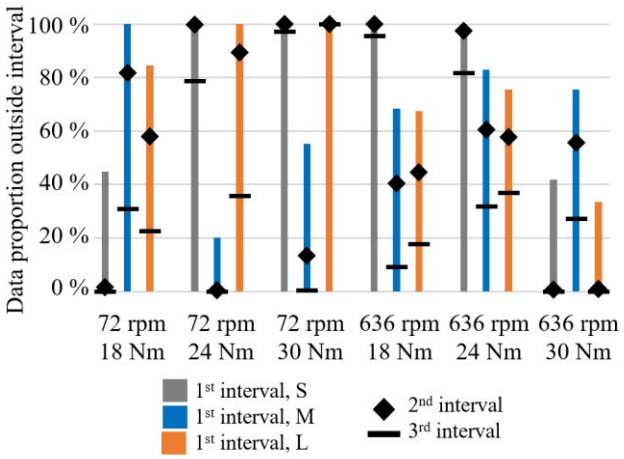


Figure 15. Encoding approach, 1st gear pair.

Figures 16 and 17 present the results of the encoding approach for the second and third gear pair in all operating conditions.

Considering the second and third gear pair, it is noticeable that the global tendency of the encoding approach corresponds to that of the first gear pair. When using the first interval, at least 39.8 % of the data for the second gear pair is always outside the limits, with one exception (28.6 % at 72 rpm and 30 Nm). For the third gear pair, a minimum of 44.9 % of the data is always outside the limits when using the first interval. If the second and third intervals are considered, the proportion of data points outside the limits decreases significantly for certain operating conditions. Especially with the third interval, some damage can no longer be detected.

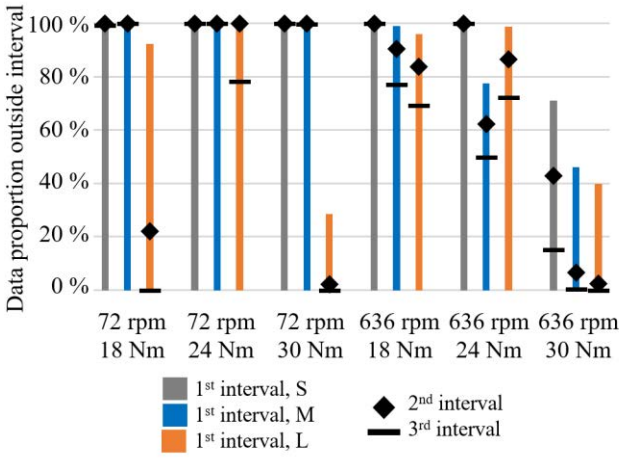


Figure 16. Encoding approach, 2nd gear pair.

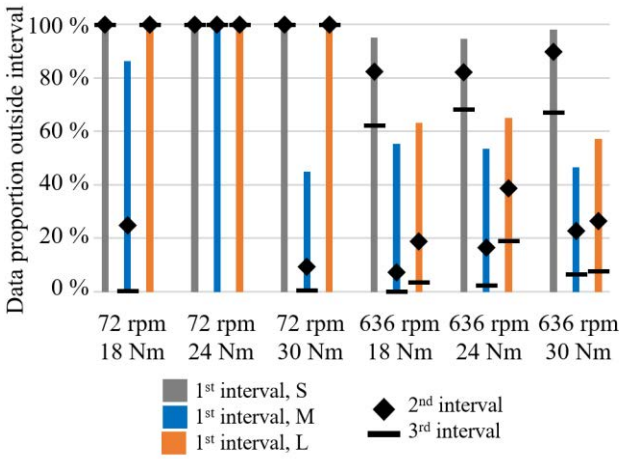


Figure 17. Encoding approach, 3rd gear pair.

3.2. Loss approach

The results of the loss approach of the first gear pair for each operating condition are given in figures 18 to 23. Again, the lower and upper limits of the first interval are determined using the data without damage (see Eq. (11)) and marked with a red line. The proportion of data points within and outside this interval is identified.

Overall, the results of the loss approach at the low speed level of 72 rpm have a low scatter of the data points. At the higher speed level of 636 rpm, larger scatter is recognizable. In all operating conditions, the data points of all pitting sizes are at least 96.8 % outside the limits – except for pitting L at operating condition 636 rpm and 18 Nm (77.0 %, see figure 21) and operating condition 636 rpm and 30 Nm (52.7 %, see figure 23). Overall, all pitting of the first gear pair can therefore be detected using the first interval of the loss approach.

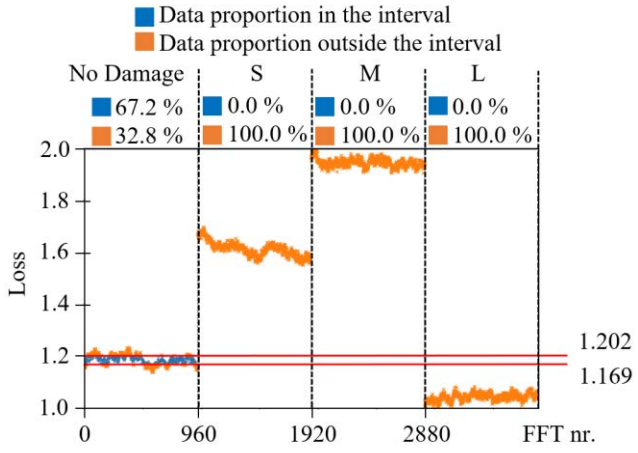


Figure 18. Loss approach, 1st gear pair, 72 rpm, 18 Nm, 1st interval.

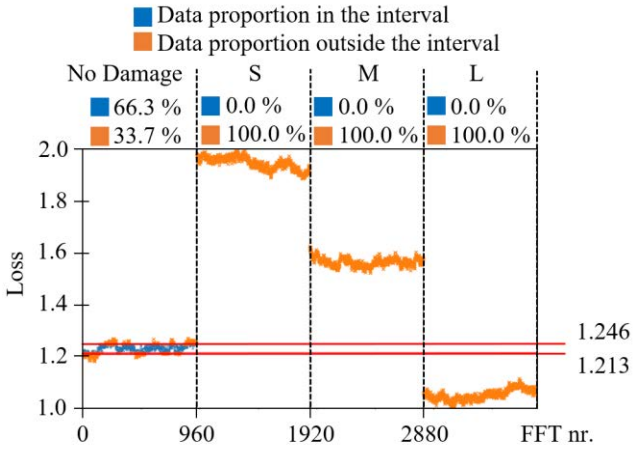


Figure 19. Loss approach, 1st gear pair, 72 rpm, 24 Nm, 1st interval.

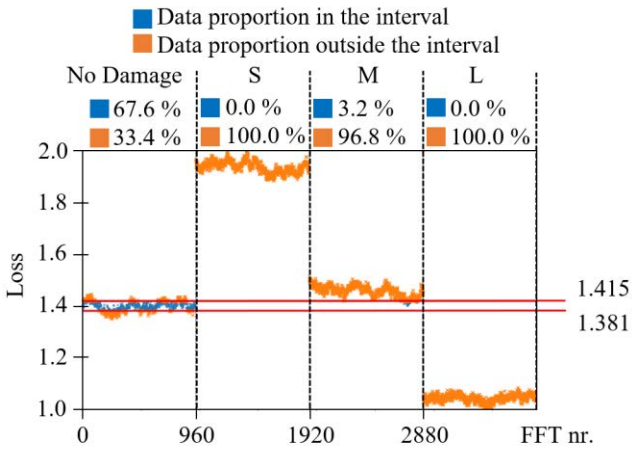


Figure 20. Loss approach, 1st gear pair, 72 rpm, 30 Nm, 1st interval.

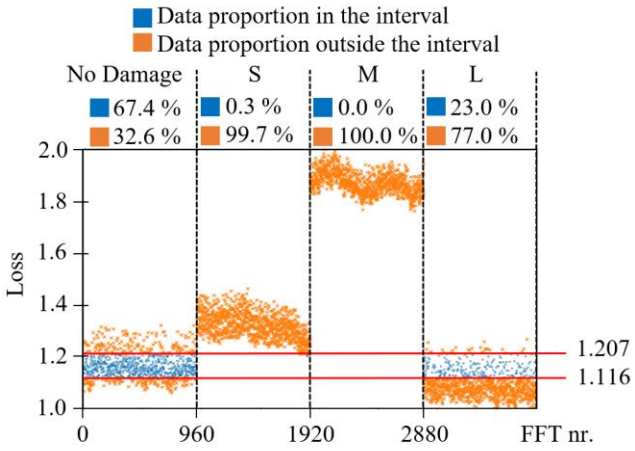


Figure 21. Loss approach, 1st gear pair, 636 rpm, 18 Nm, 1st interval.

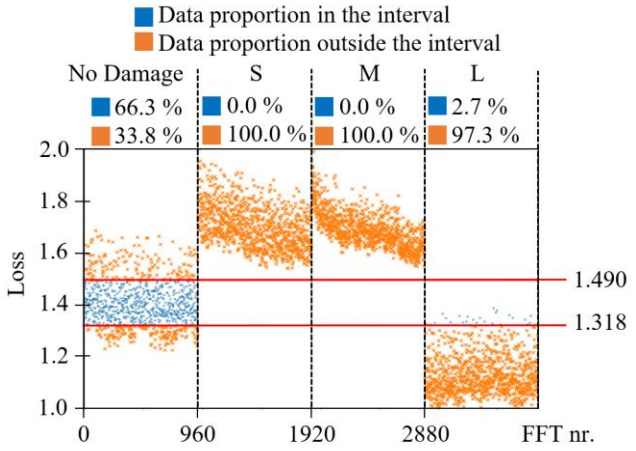


Figure 22. Loss approach, 1st gear pair, 636 rpm, 24 Nm, 1st interval.

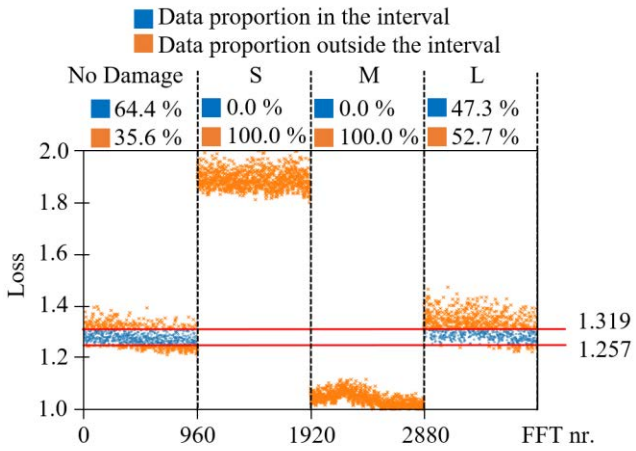


Figure 23. Loss approach, 1st gear pair, 636 rpm, 30 Nm, 1st interval.

In addition to the first interval, the second and third intervals are also determined for the loss approach (see Eq. (12) and (13)). The evaluation of all operating conditions of the first gear pair is illustrated in figure 24. The proportion of data points for damage sizes S, M and L that are outside the respective interval is shown.

When using the second interval, pitting L at operating condition 636 rpm and 30 Nm is the worst detectable pitting with only 16.4 % of the data outside the interval. Otherwise, at least 51,3 % of the data is always outside the limits. Using the third interval, pitting L is not detectable at operating condition 636 rpm and 30 Nm. Here, only 0.3 % of the data is outside the limits. Overall, pitting is more difficult to detect when using the third interval, especially at the higher speed level of 636 rpm.

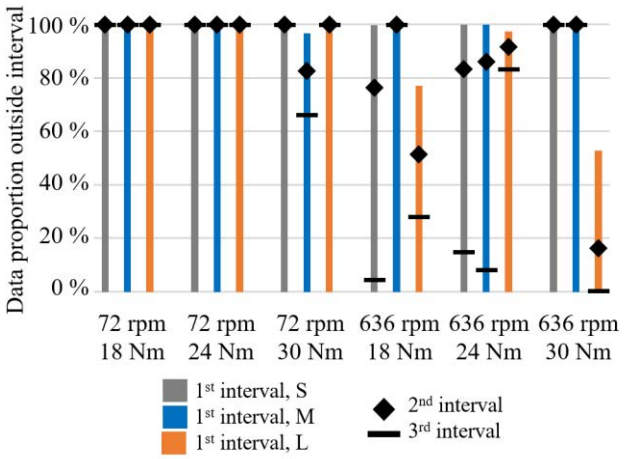


Figure 24. Loss approach, 1st gear pair.

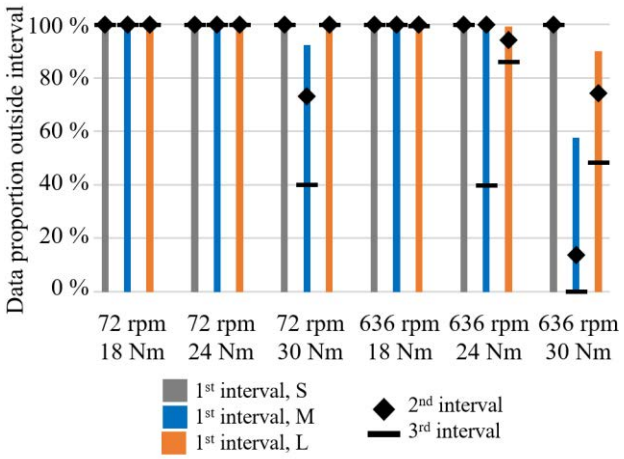


Figure 26. Loss approach, 3rd gear pair.

Figures 25 and 26 present the results of the loss approach of the second and third gear pairs for all operating conditions.

When using the first interval, at least 89.7 % of the data for the second gear pair is always outside the limits. For the third gear pair, a minimum of 57.7 % of the data is always outside the limits when using the first interval. If the second interval is calculated, 52.2 % of the data for the second gear pair is always outside the limits and 13.7 % for the third gear pair. When using the third interval, a decrease in the proportion of data points outside the limits for individual operating states and pitting sizes can be seen, similar to the first gear pair.

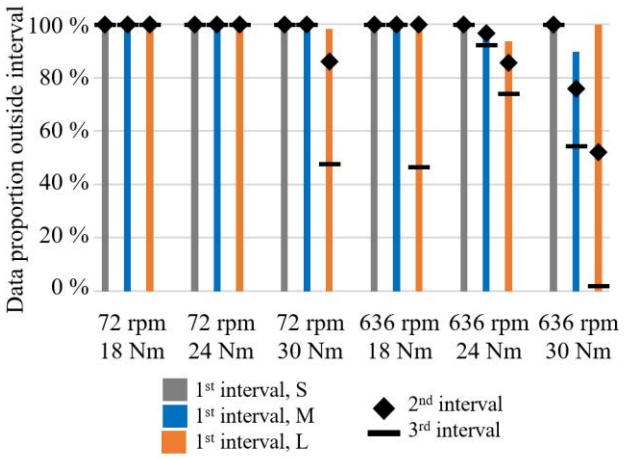


Figure 25. Loss approach, 2nd gear pair.

4. Discussion

Overall, the comparison of the encoding and loss approach shows that the data points scatter considerably more in the encoding approach. As a result, the upper and lower limits of the first interval are significantly further apart in the encoding approach than in the loss approach. When using the second and third intervals, this results in greater differences with the encoding approach because the limits are then frequently so far apart that pitting detection is no longer possible.

However, if the first interval is used for pitting detection with the encoding approach, a minimum of 20.2 % of the data points are always outside the limits for the first gear pair, 39.8 % for the second gear pair and 44.9 % for the third gear pair.

In total, the loss approach shows significantly better pitting detection than the encoding approach. When using the first interval, a minimum of 52.7 % of the data points are always outside the limits for the first gear pair, 89.7 % for the second gear pair and 57.7 % for the third gear pair.

A total of 18 cases are examined with the two approaches (3 pitting sizes in 6 operating states). In order to be able to compare the approaches even better, it is considered in how many of the 18 cases there is a significant deviation - i.e. at least 50 % of the data points outside the limits. The result can be found in figure 27.

For the first interval of the encoding approach, between 14 and 16 cases have a deviation greater than 50 %, depending on the gear pair. With the first interval of the loss approach, all damage sizes in all operating states of all gear pairs have a minimum deviation of 50 %. With the second interval of the encoding approach, the 50 % criterion only applies to 10 to 13 cases, depending on the gear pair. With the loss approach, it still applies to a minimum of 17 cases. When using the third interval, the number of cases in which at least 50 %

of the data points lie outside the interval is reduced for the encoding approach to between 5 and 12 cases. Large differences within the gear pairs can therefore also be seen here. In contrast, the loss approach shows 13 to 15 cases.

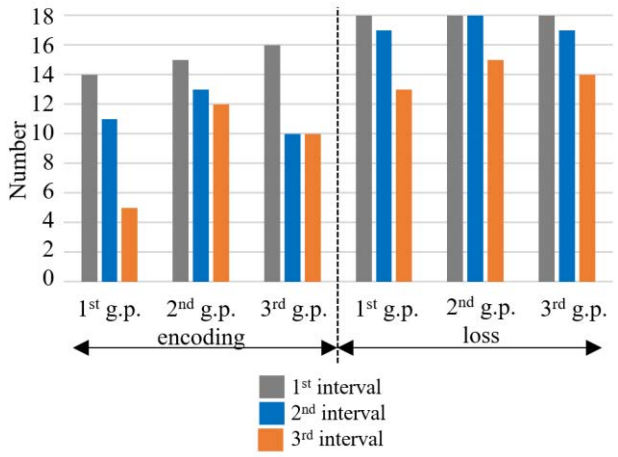


Figure 27. Number of cases with at least 50 % of the data outside the interval, per approach and gear pair (g.p.).

Overall, the loss approach is much better suitable for recognizing a clear difference between the tests without and with damage. In the second interval of the loss approach, a minimum of 50 % of the data points are still outside the limits in at least 17 cases. The encoding approach does not even achieve this for the first interval. In addition, the loss approach is not only less sensitive to different damage sizes and operating conditions, but also to different gear pairs. However, this paper does not consider how well false positives can be excluded with the two approaches.

Regardless of the comparison of the two approaches, even the smallest investigated pitting with a size of 0.61 % (1st gear pair), 0.51 % (2nd gear pair) and 0.78 % (3rd gear pair) could be detected in the context of this study. In contrast to the approach presented in (Binanzer et al., 2023), in which an AE was combined with a Long Short Term Memory (LSTM) network, detection is also possible with a purely unsupervised algorithm. This offers the advantage for the application that no labeled training data is required. Only data from a test without damage is required for training.

The detectable pitting sizes in the scope of this work are a significant improvement on other investigations. There are various approaches for pitting detection in gearboxes using vibration sensors. The approaches differ on the one hand in the investigated pitting size and in the methods of sensor data evaluation.

Qu, M. He, Deutsch and D. He (2017) investigated one row of pitting damage along the tooth width of one tooth. A stacked autoencoder network was used to perform the dictionary learning in sparse coding and automatically extract features from the raw vibration data. With these features a backpropagation neural network was trained to identify the damage.

Fan, Zhou, Wu and Guo (2017) developed a gear damage detection and localization approach by analyzing the vibration signal of an individual tooth and Support Vector Machines (SVM). The dispersion degree and vibration accelerations of the waveform of an individual gear tooth were studied to investigate the characteristics of gear tooth under normal, small failure (< 5 % damaged tooth area) and serious failure (> 5 % damaged tooth area) conditions.

An unsupervised feature extraction method called disentangled tone mining was presented by Qu, Zhang, M. He, D. He, Jiao and Zhou (2019). This method was able to identify the fault level directly from the frequency spectrum of the measured vibration data. Pitting sizes between 4.33 % and 24.91 % were investigated in a single stage spur gearbox.

Medina, Cerrada, Cabrera, Sanchez, Li and Oliveira (2019) used a LSTM network for classifying nine levels of pitting. The smallest investigated pitting had a size of 4.16 %.

Pitting sizes of less than 1 % were detected by Grzeszkowski, Nowoisky, S., Scholzen, Kappmeyer, Gühmann, Brimmers and Brecher (2020) using a SVM classifier. A disadvantage of the SVM classifier is that it is a supervised algorithm and therefore requires labeled training data.

Damage detection with purely physically based data evaluation, with pitting sizes between 6.3 % and 41.7 %, was presented by Sowana und Chandrasekaran (2020). In each case, the root mean square (RMS) value of the structure-borne noise data in the time domain of the undamaged and damaged gear was compared.

Sarvestani, Rezaeizadeh, Jomehzadeh and Bigani (2020) also examined the detection of naturally occurring pitting damages with a size of 30 %, 60 % and 90 % using purely physically based methods. The frequency spectrum of the structure-borne noise data was divided into six ranges. The damage was best detected in the second gear mesh harmonic range.

Häderle, Merkle and Dazer (2024) presented another physically based data analysis approach. It is shown that the greatest percentage difference between undamaged and damaged gears can be determined for the harmonics of the gear mesh frequency (GMF) and the sidebands between 24,000 Hz and 40,300 Hz. Thus, it was possible to detect very small pitting sizes between 0.42 % and 1.83 %.

5. Conclusion

In order to increase the service life of gearboxes, avoid unexpected failures and thus reduce overall operating, maintenance and labor costs, comprehensive PHM has to be implemented in gearbox applications. Adaptive operating strategies can even extend the RUL without any loss of performance. In order for the PHM of gearboxes to achieve its full

potential, damage detection at the earliest possible stage is essential.

In this study, two unsupervised machine learning approaches (encoding and loss approach) were developed and the detection of artificially manufactured damage on the tooth flank of a test gearbox was investigated.

In particular, the loss approach is more capable of identifying a difference between no damage and damage than the coding approach, regardless of the size of the pits and operating conditions. The loss approach is also less sensitive to different gear pairs, which have slightly different properties due to material and manufacturing tolerances.

Overall, it can be stated that the main contribution of this work is that existing Machine Learning tools have been applied to the challenge of a very early damage detection in gearboxes. Without the need of complex physically based evaluation methods of the vibration data, the smallest pitting of about 0.5 % could be detected regardless of the operating condition. The use of the sparse AE was described in detail and two evaluation methods were compared.

REFERENCES

Bertsche, B. & Dazer, M. (2022). *Zuverlässigkeit im Fahrzeug- und Maschinenbau: Ermittlung von Bauteil- und System-Zuverlässigkeiten.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Binanzer, L., Merkle, L., Dazer, M. & Nicola, A. (2023). Pitting Detection for Prognostics and Health Management in Gearbox Applications. *International Conference on Gears 2023* (*VDI-Berichte*, vol. 2422, pp. 97-108), September 13-15, Munich, Germany. doi:10.51202/9783181024225

Fan, Q., Zhou, Q., Wu, C. & Guo, M. (2017). Gear tooth surface damage diagnosis based on analyzing the vibration signal of an individual gear tooth. *Advances in Mechanical Engineering (AIME),* vol. 9 (no. 6), pp. 1-14. doi: 10.1177/1687814017704356

German Institute for Standardization (DIN) (2010). Industrial liquid lubricants - ISO viscosity classification (ISO 3448:1992). In DIN, *DIN ISO 3448:2010-02*. Berlin, Germany: Beuth Verlag GmbH. doi: 10.31030/1562009

Goebel, K., Celaya, J., Sankararaman, S., Roychoudhury, I., Daigle, M. & Abhinav, S. (2017) Prognostics: The Science of Making Predictions. CreateSpace Independent Publishing Platform.

Gretzinger, Y., Lucan, K., Stoll, C., & Bertsche, B. (2020). Lifetime Extension of Gear Wheels using an Adaptive Operating Strategy. *Proceedings of 7th International Conference Integrity-Reliability-Failure* (pp. 703-710), September 6-10, Funchal, Portugal. https://fe.up.pt/irf/Proceedings_IRF2020/

Grzeszkowski, M., Nowoisky, S., Scholzen, P., Kappmeyer, G., Gühmann, C., Brimmers, J. & Brecher, C. (2020). Classification of Gear Pitting Severity Levels using Vibration Measurements. *In tm - Technisches Messen*, vol. 87 (no. s1), pp. s56-s61. doi: 10.1515/teme-2020-0026

Häderle, P., Merkle, L. & Dazer, M. (2024). Vibration Analysis for Early Pitting Detection During Operation. *Forschung im Ingenieurwesen.* vol. 88 (article nr. 15). doi: 10.1007/s10010-024-00743-5

Institute of Electrical and Electronics Engineers (IEEE) (2017). IEEE Standard Framework for Prognostics and Health Management of Electronic Systems. In IEEE, *IEEE Std 1856-2017* (pp. 1-31). doi: 10.1109/IEEESTD.2017.8227036

International Organization for Standardization (ISO) (2016). Calculation of load capacity of spur and helical gears – Part 5: Strength and quality of materials. In ISO, *ISO6336-5:2016(E)*, (p. 5). Genève, Switzerland: International Standards Organization.

Medina, R., Cerrada, M., Cabrera, D., Sanchez, R.-V., Li, C. & Oliveira, J. V. D. (2019). Deep Learning-Based Gear Pitting Severity Assessment Using Acoustic Emission, Vibration and Currents Signals. *Proceedings of 2019 Prognostics and System Health Management Conference (PHM-Paris 2019)* (pp. 210-216), May 2-5, Paris, France. doi: 10.1109/PHM-Paris.2019.00042

Ng, A. (2011). Sparse autoencoder. CS294A Lecture notes, 72, pp. 1-19.

Qu, Y., He, M., Deutsch, J. & He, D. (2017). Detection of Pitting in Gears Using a Deep Sparse Autoencoder. *Applied Science. Special Issue Deep Learning Based Machine Fault Diagnosis and Prognosis*, vol. 7 (no. 5), pp. 515-529. doi: 10.3390/app7050515

Qu, Y., Zhang, Y., He, M., He, D., Jiao, C. & Zhou, Z. (2019). Gear pitting fault diagnosis using disentangled features from unsupervised deep learning. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability,* vol. 233 (no. 5), pp. 719-730. doi: 10.1177/1748006X18822447

Research Association for Drive Technology (FVA) (1985). FVA-Heft 180 Referenzöle - Datensammlung für Mineralöle.

Sarvestani, E. S., Rezaeizadeh, M., Jomehzadeh, E. & Bigani, A. (2020). Early Detection of Industrial-Scale Gear Tooth Surface Pitting Using Vibration Analysis. *Journal of Failure Analysis and Preventionm,* vol. 20, pp. 768-788. doi: 10.1007/s11668-020-00874-1

Sonawane, P. R. & Chandrasekaran, M. (2020). Investigation of gear pitting defect using vibration characteristics in a single-stage gearbox. *The International Journal of Electrical Engineering & Education*, vol. 57 (no. 3). doi: 10.1177/0020720918813837

Data Scarcity in Fault Detection for Solar Tracking Systems: the Power of Physics-Informed Artificial Intelligence

Mila Francesca Lüscher[1], Jannik Zgraggen[2], Yuyan Guo[3], Antonio Notaristefano[4], and Lilach Goren Huber[5]

[1,2,5] *Zurich University of Applied Sciences, Technikumstrasse 81, Winterthur Switzerland*
mila.luescher@zhaw.ch
jannik.zgraggen@zhaw.ch
lilach.gorenhuber@zhaw.ch

[2] *Fluence Energy LLC, Hornbachstrasse 50, CH-8008 Zurich, Switzerland*
yuyan.guo@fluenceenergy.com
antonio.notaristefano@fluenceenergy.com

ABSTRACT

Combining physical and domain knowledge in artificial intelligence (AI) models has been gaining attention in various fields and applications. Applications in machine prognostics and health management (PHM) are natural candidates for such hybrid approaches. In particular, they can be efficiently exploited for high fidelity anomaly detection in technical and industrial systems. A natural way for hybridization is using physical models to generate representative data for the training of AI models. Depending on the level of domain knowledge availability, data augmentation can compensate for scarcity of real data from the field. This is particularly attractive for anomaly detection tasks, in which data from the abnormal regimes is limited by definition. On top of this inherent data limitation, many real-world systems suffer from data limitations even within the normal regimes.

In this paper we propose a physics-informed deep learning algorithm for fault detection in grid scale photovoltaic power plants. We focus on a common data scarce scenario that emerges from a low asset monitoring granularity: instead of monitoring the power production of each solar string, the power output is monitored only at combiner-box or even inverter level (monitoring a large number of strings with a single sensor). As a result, the signatures of single local faults can become very subtle and challenging to detect. We show that in this case a physics-informed AI approach significantly outperforms the alternative of a purely data-driven anomaly detection model. This enables high fidelity fault detection in larger solar power plants, that are often limited in the granularity of their condition monitoring data.

1. INTRODUCTION

Utilizing physical information and domain knowledge in conjunction with AI models has become a popular approach to deal with some of the known limitations of AI (Karniadakis et al., 2021), such as the lack of interpretability of AI models and their data-hungry nature. The field of equipment prognostics and health management (PHM) is an ideal application field for such hybrid approaches (Rausch, Goebel, Eklund, & Brunell, 2005; Wu, Sicard, & Gadsden, 2024). For many of the systems, a detailed physical model is already in use for design purposes (Chao, Kulkarni, Goebel, & Fink, 2019; Huber, Palmé, & Chao, 2023), and can be exploited also for PHM. In other systems the fault or degradation mechanisms are well understood and allow for a microscopic or a phenomenological model (Rai & Mitra, 2021; Zgraggen, Guo, Notaristefano, & Goren Huber, 2023).

A typical challenge in PHM tasks is the severe lack of historical failure data. In these cases, the use of physical information to compensate for data scarcity becomes even more attractive than in other application domains. One particularly common approach is to augment the training data using physical models (Frank et al., 2016; Wu et al., 2024). Such models can be used either for operational regimes that are scarce on data (Chao, Kulkarni, Goebel, & Fink, 2022; W. Li et al., 2021), or to directly model fault mechanisms that are rarely seen in operation (Kohtz, Xu, Zheng, & Wang, 2022; Bansal et al., 2022).

In our previous work we took the latter approach (Zgraggen, Guo, Notaristefano, & Goren Huber, 2022). We developed a physical model that corrupts data from a normally operating photovoltaic (PV) plant, thereby generating data with syn-

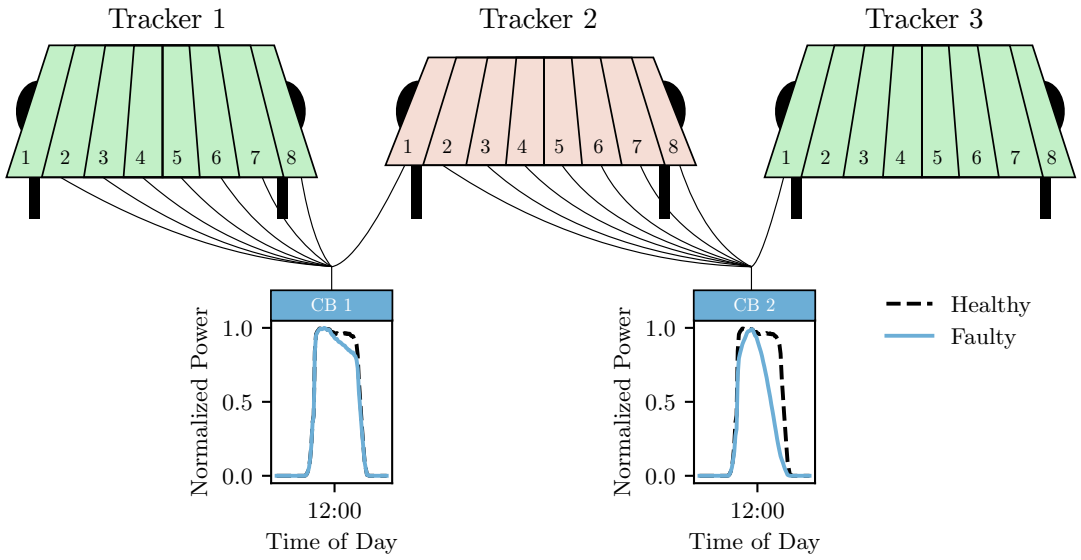


Figure 1. The challenge of low data granularity for tracker fault detection. The eight strings that are mounted on the faulty tracker 2 contribute their power to two different combiner boxes (CBs): CB1 is affected by only one of these strings. Its power profile (bottom left, solid blue) suffers a mild loss compared to the daily reference (dashed black). CB2 is affected by seven of the faulty strings, and suffers a more significant power loss (bottom right).

thetic faults. In this paper we extend this method to allow its applicability under data scarcity in real-world scenarios. The proposed hybrid approach profits from both worlds: on one hand it demonstrates a high ability to mimic the effects of rare faults, without the need for real faults in the data. On the other hand, it does not require a complex physical model of the normal system, as all complex (environmental and operational) effects are already captured by the field data. As opposed to previous approaches to solar plant fault detection, our method is independent of lab data (Chen, Chen, Wu, Cheng, & Lin, 2019; B. Li, Delpha, Diallo, & Migan-Dubois, 2021; Gao & Wai, 2020), simulation data (Chine et al., 2016), or designated data-collecting hardware (Daliento et al., 2017; Amaral, Pires, & Pires, 2021), and was carried out using existing operational data only. Our Physics-Informed Deep Learning (PIDL) approach was shown to perform very accurately with no need for fault data (Zgraggen et al., 2022), and even in a fully unsupervised setting, where the data may be contaminated by unlabeled anomalies (Zgraggen et al., 2023). Moreover, the approach does not require any irradiance measurements, but merely the standard 15-minute measurements of the power output from individual PV strings. Also in this respect our work is rather unique: most of the published work related to PV plant fault detection (Mellit, Tina, & Kalogirou, 2018; Triki-Lahiani, Abdelghani, & Slama-Belkhodja, 2018; Pillai & Rajasekar, 2018; Mansouri, Trabelsi, Nounou, & Nounou, 2021) relies on data at single module or cell resolution, rather than the operationally relevant string-data, often containing dozens or hundreds of modules.

In grid-scale solar power plants, it is often impractical to monitor data at string level due to the large number of PV strings involved. As a result, individually monitoring each string often becomes unfeasible. In this case, the output power is monitored and recorded only at a higher spatial granularity level, for example at the level of combiner boxes or even inverters, gathering a large number of strings in a single sensor reading. As shown below, this lower monitoring granularity inevitably leads to a reduced effectiveness in detecting local faults. To the best of our knowledge, there are no previously published studies that address fault detection at combiner-box or inverter level in PV power plants.

In this paper we address the above common scenario of low data granularity by extending our previous PIDL approach. We use a physical model to transfer the method from assets with a high data granularity to assets with a low data granularity. We show that in the case of data scarce assets, the physics-informed (PI) approach is of an even higher benefit compared to purely data-driven anomaly detection.

The contribution of this paper is two-fold. For solar power plant condition monitoring, it offers a high fidelity method to detect anomalous power losses by combining physical knowledge and AI in real-world operational conditions. In a more general context, the paper demonstrates the effectiveness of physics-informed AI for fault detection in data-scarce scenarios, which are common in various application fields. In particular, we show that physical knowledge can be utilized for transfer learning between domains with abundant data and domains with scarce data.

In Section 2 we describe the solar tracker use-case on which

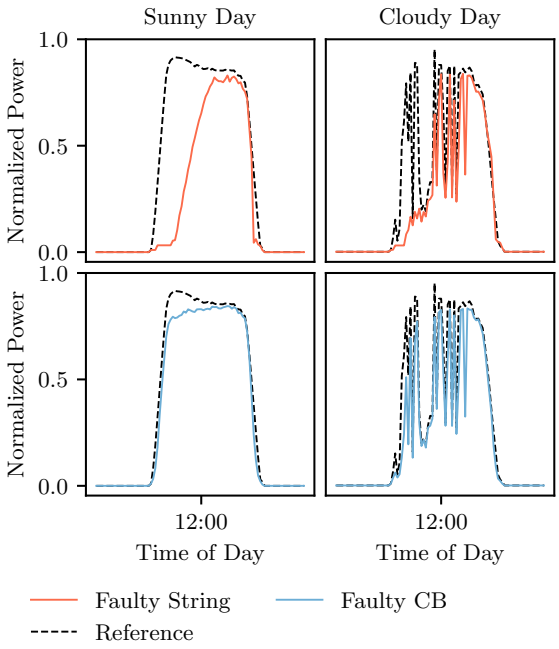


Figure 2. The effect of low data granularity on the measured output power. The signature of a real tracker fault on the daily output power of a single string (upper panels) and of the entire combiner box (CB, lower panels). A clear-sky day (left panels) is contrasted with a cloudy day (right panels). The fault signatures are considerably smaller and harder to detect, if only CB-level power data is available.

we demonstrate our approach. In Section 3 we provide the details of the PIDL approach. Finally, the results are shown and discussed in Section 4.

2. Description of the Use Case

The proposed PIDL approach is demonstrated here for the early detection of faults in the tracking system of solar power plants. Solar trackers are rotating units on which PV panels are mounted in order to adjust their orientation during the day according to the position of the sun, thus ensuring maximal power production at any given moment (Racharla & Rajan, 2017). In a common fault mechanism of solar trackers, the trackers get stuck at a certain orientation instead of following the sun. This fault has an immediate implication on the power production, which is significantly reduced compared to the optimum, given certain irradiance and weather conditions. Thus, an automatic early detection of the fault by closely monitoring the power production patterns can significantly reduce the resulting energy losses.

In our previous work (Zgraggen et al., 2022) we developed an algorithm for early detection of tracker faults based on power profiles of PV strings. The algorithm is thus applicable to power plants in which the power production is monitored for each PV string individually. However, a large fraction of the operational PV power plants nowadays are monitored at a lower granularity, that is, at the level of combiner boxes (CB) or even inverters. In such cases, historical power data is only available for single CBs or inverters, summing up the power of up to tens of individual strings. The single string power is no longer available, thus the previously proposed fault detection algorithm is not directly applicable.

To understand the fault detection challenge posed by the lower data granularity, an example is illustrated in Figure 1, showing two CBs with their related trackers. Since a CB extends over a large area, its strings are typically mounted on several different solar trackers, in this case trackers 1,2 and 3. Thus, if one tracker is faulty, only a fraction of the CB power originates from a string that is affected by the fault while the rest of the strings of this CB do not display any signatures of the tracker fault. In the illustration of Figure 1, Tracker 2 is faulty, while Trackers 1 and 3 are normally functioning. Combiner box CB1 receives its input from 7 strings which are unaffected by the tracker fault (as they are mounted on Tracker 1) and one string which is affected by the fault (as it is mounted on Tracker 2). As a results, the CB power profile (shown at the bottom left in blue) is only mildly impacted by the fault, compared to the reference profile (dashed black). On the other hand, CB2 receives its input from 7 affected strings (mounted on Tracker 2) and only one unaffected string (on Tracker 3). The resulting CB2 power profile (bottom right in blue) shows a much stronger fault signature than the one of CB1. Note that the black dashed profiles are the daily reference power production, calculated from the entire plant data (see explanation in Sec. 3). Moreover, it should be noted that the example is illustrated for a sunny day with clear sky, whereas the effectiveness of the proposed method is shown below under any weather and operational conditions.

As argued above, typical fault signatures on CB power profiles are much more subtle than on string power profiles, and require a higher anomaly detection sensitivity to identify and locate them. Figure 2 demonstrate this effect using data from a real operational PV plant, under different weather conditions. The signatures of a tracker fault on the measured output power are shown at the two monitoring levels: string level vs. CB level. In the upper panels we display (normalized) daily power profiles of a single string which was mounted on a faulty tracker, compared to the daily reference (dashed black). In the lower panels we assume that string level data is unavailable and display the power profiles of the entire CB containing the same string of the upper panels. Since this CB sums up the power of both faulty and intact strings, the signature of the tracker fault is smaller and harder to detect. This is particularly true under cloudy weather conditions, as shown at the right column.

The focus of this paper is the transfer of the tracker fault

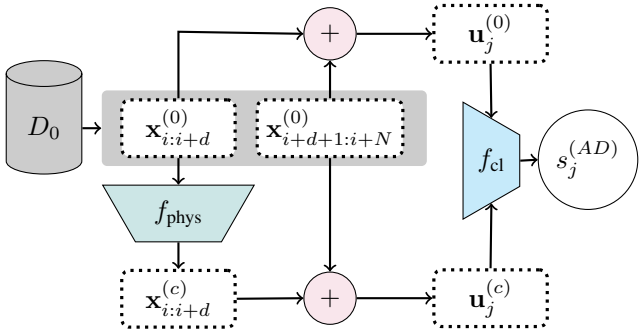


Figure 3. The proposed Physics-Informed AI fault detection algorithm.

detection algorithm from string level monitoring data to CB level monitoring data, thereby addressing the high fault detection sensitivity challenge.

3. METHOD

In order to achieve high fidelity fault detection of tracker faults for CB-level monitoring data we introduce an extension of our previous string-level PIDL model. The proposed PIDL algorithm for tracker fault detection includes two steps: (i) Data augmentation using a physical model that synthetically generates abnormal power profiles. (ii) Training a binary classifier to distinguish normal from abnormal daily power profiles.

3.1. Physics Informed Data Augmentation

Due to the rare occurrence of tracker faults, real operational power data which is affected by these faults is very scarce. However, data from normal functioning solar power systems is abundant. We exploit this fact, and use operational power data from normally functioning solar plants in order to generate synthetic power profiles under tracker faults. Since the tracker fault mechanism is well understood, we develop physical equations that enable a simple transformation of a healthy power profile into a faulty one. In this way we can simulate diverse fault scenarios and augment the training data with a large number of realistic tracker fault examples. In a second step, the augmented data containing both healthy and faulty power profiles is used to train a binary classifier that distinguishes between normal and abnormal power profiles, thereby enabling identification and localization of tracker faults in large power plants. In the following we describe the physics informed data augmentation method.

A tracker fault affects the power production of the solar strings that are mounted on the faulty tracker. Neighboring strings, if mounted on healthy functioning trackers, remain unaffected. In particular, a common situation (as illustrated in Figure 1) is that out of the N strings that are combined into one CB, only $d < N$ are mounted on a faulty tracker and the rest $N - d$ strings are mounted on healthy trackers.

In order to synthetically generate CB power profiles that correspond to various types of tracker faults, we model d faulty string power profiles that result from a tracker getting stuck at an angle θ_0. This is done using a physical model f_{phys} of the fault mechanism that "corrupts" normal power profiles of single strings, turning them into faulty profiles. The d synthetically generated faulty profiles are added to $N - d$ real healthy string profiles from the operational system, to obtain a synthetic CB profile which is partially affected by a tracker fault, as depicted in Figure 3.

The generation of a faulty string power profile $x^{(c)}(t)$ out of a healthy string profile $x^{(0)}(t)$ is done using the equations

$$\begin{aligned} x^{(c)}(t) &= c_p \left[(1-\gamma) g(\theta_0, \theta^*_{\text{i}}(t)) + \gamma\right] x^{(0)}(t) \\ g(\theta_0, \theta^*_{\text{i}}(t)) &= \frac{\cos\theta_0 \cdot f_{\text{IAM}}(\theta_0)}{\cos\theta^*_{\text{i}}(t) \cdot f_{\text{IAM}}(\theta^*_{\text{i}}(t))}. \end{aligned} \tag{1}$$

with $f_{\text{IAM}}(\theta_{\text{i}}) = 1 - b_0(1/\cos\theta_{\text{i}} - 1)$ and where $\theta^*_{\text{i}}(t)$ is the optimal tilt angle of the tracker at time t, θ_0 is the stuck angle of the faulty tracker, b_0 and γ are model parameters estimated empirically using the data, by fitting 10 samples of faulty profiles from the operational data of the string-level PV plant (we note that such profiles are only needed for a single plant, and are not required for the target plant at CB level). The parameter c_p is a degradation loss coefficient, assumed to range between 0.8 and 1 in order to simulate slight losses which are unrelated to tracker faults, and may exist also in healthy strings. For details of the physical model we refer the reader to (Zgraggen et al., 2022).

By adding up d faulty and $N - d$ normal string profiles, a synthetically generated faulty CB power profile $u^{(c)}(t)$ obtains the form

$$u^{(c)}(t) = \frac{1}{d}\sum_{i=1}^{d} x_i^{(c)}(\theta_0, \gamma, b_0, c_p; t) + \frac{1}{N-d}\sum_{i=d+1}^{N} x_i^{(0)}(t) \tag{2}$$

where $x_i^{(c)}(\theta_0, \gamma, b_0, c_p; t)$ is the ith corrupted string profile and $x_i^{(0)}(t)$ is the ith healthy string profile. The model parameters θ_0, γ, b_0 and c_p are sampled from uniform distributions within realistic ranges to represent all physically viable configurations (see (Zgraggen et al., 2022) for details), but are kept identical for all of the strings that belong to the same CB. The number of corrupted strings d is drawn randomly from the range $1...N$ in order to cover all possible configurations under the constraint of N strings in one CB (which is given by the plant configuration).

In addition to the generation of faulty CB profiles, we generate an equal amount of healthy CB profiles by simply adding N healthy adjacent string profiles. Note that we follow the modelling approach described in (Zgraggen et al., 2022), in which we randomly introduce mild physics-informed modifications to the healthy profiles in order to mimic the ef-

fects of small power losses that are unrelated to tracker faults. As shown in (Zgraggen et al., 2022), allowing this physics-inspired stochastic variability in the training data, increases both the accuracy and the robustness of the model predictions.

The proposed data augmentation process ensures a large diversity of tracker faults with different intensities, stuck angles and under various soiling or degradation conditions. Moreover, an important advantage of our approach is its mathematical structure, that enables using real operational power profiles and transforming them into faulty profiles by mathematically "injecting" a known fault mechanism into them. As a result, complex features of the model inputs, such as diverse weather effects, are already accounted for, and do not need to be modeled.

We note that the CB-level model described above uses string power profiles to generate CB power profiles. As such, it assumes the availability of normal data from one power plant which is monitored at string level. The results we show below were obtained after training on data from a string-level plant (the source plant), but tested on an operational power plant with CB monitoring only, in a different geographic location (the target plant). With this we demonstrate that effective fault detection is transferable to the target plant without string-level data availability, owing to the physics informed modeling approach.

3.2. CNN fault classifier

The empirical-physical model of the fault mechanism is used to augment the normal data set, such that it now contains healthy as well as faulty power profiles at CB level, $u_j^{(0)}(t)$ and $u_j^{(c)}(t)$ respectively. Each daily profile is a time-series of size 96 (due to a 15 minute resolution of the original sensor data). At a next step, the augmented data set, containing balanced healthy and faulty samples is pre-processed by subtracting from each power profile the daily reference profile, calculated as the 0.9 quantile over the entire plant at any given moment in time (see (Zgraggen et al., 2022) for details). The resulting power deviation profiles are used to train a 1d-CNN classifier f_{cl} that assigns an anomaly score $s_j^{(\text{AD})}$ to each daily profile, as depicted in Figure 3. This allows to detect faulty combiner boxes (thereby locating the related faulty trackers) at the end of each day, which is the relevant time resolution for decision making in practice. The CNN contains three one-dimensional convolutional layers followed by two fully-connected layers, with a total of around 30'000 trainable parameters. The network architecture was optimized using a grid search to tune the number of layers and filters and the learning rate.

We trained the classifier with 700'000 CB power profiles, half of which include synthetic tracker fault effects. All profiles originate from one single PV power plant during a time period of one year. The test data originates from another PV power plant, monitored at CB level, and includes 5349 CB power profiles collected during a time period of two months and containing 857 known faulty profiles, labeled manually by domain experts.

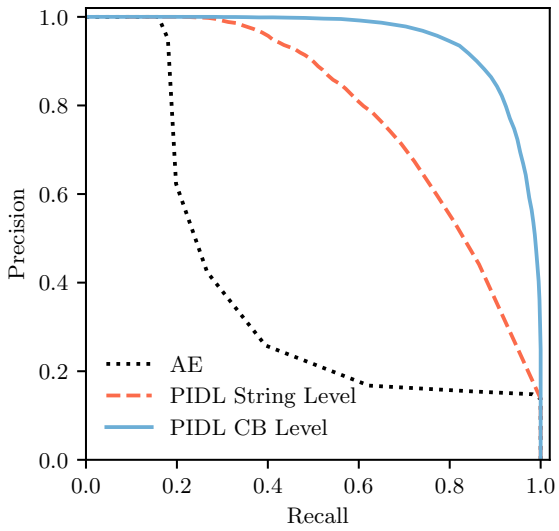


Figure 4. Fault detection evaluation using precision-recall curves. The performance of the proposed CB-level PIDL model (solid blue) is compared with a simpler string-level PIDL (dashed red) and a purely data-driven convolutional AE model (dotted black).

Baselines. We compare the performance of the proposed algorithm with two baseline methods. The first one is a similar PIDL algorithm which is trained using the original string level profiles, rather than CB-level profiles, with and without synthetic faults. This enables us to examine the transferability of the learned features from string to CB level.

The second baseline we compare to is a purely data-driven approach, not making use of any physics-based modeling. In this case we train a convolutional Autoencoder (AE) neural network to reconstruct power profiles. The AE is trained with the normal part of the data only, not including any tracker faults. The normalized reconstruction errors are then used as fault indicators, with a threshold typically set at the tail of the training distribution of reconstruction errors. The feature extraction layers of the AE are four 1d-convolutional layers, similarly to the PIDL network described above, with a similar number of 46'000 trainable parameters.

4. Results

The performance of the proposed PIDL classifier is evaluated in Figure 4 using a precision-recall curve (PRC). The PRC of the CB-level PIDL method is shown in solid blue, and is

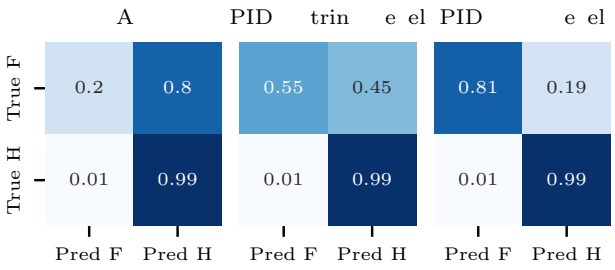


Figure 5. Fault detection evaluation using confusion matrices. The performance of the proposed CB-level PIDL model (right) is compared to the ones of a string-level PIDL (middle) and a purely data-driven AE (left). For all 3 models, the detection thresholds were set to yield a false positive rate of 1%.

contrasted with the PRC of the string-level PIDL approach (dashed red), and the PRC of the pure data-driven AE model (dotted black). It is evident that a pure data-driven approach does not exploit the physical knowledge of the fault mechanism and thus reaches a much poorer performance than both of the PI approaches. In the case of string-level PIDL, the inputs are not aggregated to mimic CB power profiles whereas in the case of CB-level PIDL the string power profiles are aggregated in various ways such that different tracker fault configurations are simulated in the synthetic data. In this way, also configurations that lead to subtle fault signatures are introduced at training, and can be detected at inference time. As seen from the PRC results, this leads to a significant improvement in the fault detection performance, with an average precision (AP) of 0.95 for the proposed CB-level PIDL, compared to 0.79 for the string-level PIDL. As expected, the purely data-driven AE model is significantly inferior in its fault detection performance, with an AP of 0.38. It should also be noted that the performance of the CB-level PIDL is only slightly worse than the one we reported for the string-level PIDL when tested on string-level data (with an AP of 0.97, see (Zgraggen et al., 2022)).

In addition to the PRC, we compare the fault detection performance using confusion matrices shown in Figure 5. The confusion matrix of the CB-level PIDL model (right) is compared with the ones of the string-level PIDL (middle) and the data-driven AE (left). For the sake of model comparison, all three confusion matrices were generated by selecting a detection threshold that guarantees a low false positive rate of 1%. This is a practically sensible threshold, that reduces the false alarms to a minimum. Fixing the threshold to produce this false positive rate on the test data in all three methods, we obtain a false negative (missed detections) rate of 0.8 with the pure data-driven approach, a rate of 0.45 with the string-level PIDL and a significantly lower rate of 0.19 with the proposed CB-level PIDL algorithm.

The complexity of the fault detection task is demonstrated in Figure 6 using CB power test data from the target power plant. Each panel displays a CB daily power profile (solid blue) compared with the daily reference (dashed black). The upper six panels are examples of CB power profiles with no tracker faults, whereas the six lower panels were labeled as suffering from power losses due to tracker faults. The power profiles in the 6 panels at the left half of the figure were all correctly classified by the proposed PIDL algorithm, as well as by the purely data-driven convolutional AE. Indeed, the fault signatures of the three profiles at the bottom left are rather strong and could be clearly assigned to tracker faults by both models. This stands in contrast to the 6 panels on the right hand side of the figure, which were all correctly classified by the PIDL model, but misclassified by the AE. Here, physical information about the tracker fault mechanism clearly helped to distinguish between true tracker faults (lower panels) and power losses due to other reasons, unrelated to the solar trackers (upper panels). This is despite the fact that such unrelated power losses may be rather high, as seen in the three upper right panels. In all three cases, due to their high power losses compared to the reference, the AE produced high reconstruction errors, leading to false positives. On the other hand, the low power losses of the truly faulty profiles at the bottom right led to missed detections (false negatives) by the AE, because of reconstruction errors that are similar in magnitude to the ones of the training data. Despite their low power losses, and their mild fault signatures, these power profiles were correctly detected as suffering from tracker faults by the CB-level PIDL algorithm.

To conclude, the CB-level PIDL includes a physics-informed data augmentation step that captures important nuances in the fault features, even in case of low data availability that leads to very mild fault signatures. The same data augmentation framework can be easily generalized to any monitoring level, provided the structure of the monitoring data at the operational plant (i.e number of strings per combiner-box or inverter). The only prerequisite is the availability of string level power profiles from a normal functioning power plant that can serve as the baseline for data augmentation. Moreover, one of the advantages of our approach is that it does not require complex measurement and/or modeling of the solar irradiance under various ambient conditions, but relies entirely on a single measured variable: the output power.

The proposed approach of physics-informed data augmentation is generally applicable in systems with some understanding of the fault mechanism. However, we believe that this physical understanding does not need to be complete or to amount to a full microscopic model of the fault mechanism. In many cases, a phenomenological model of the fault signatures on the observed data may be sufficient in order to achieve superior fault detection performance compared to purely data-driven approaches.

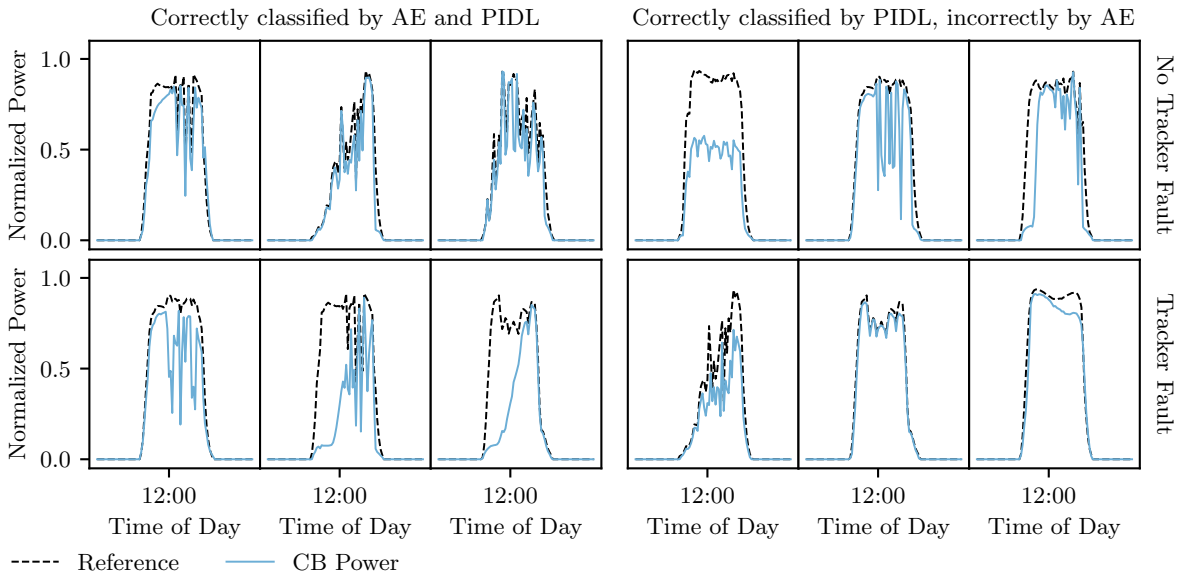


Figure 6. Classification outcomes of the PIDL compared with a purely data-driven AE model. Each panel displays a CB power profile (solid blue) together with the daily reference profile (dashed black). True labeled profiles with tracker faults (bottom row) are contrasted with profiles with no tracker faults (top row). The 6 panels on the left half were correctly classified by both the PIDL and the AE models, whereas the 6 panels on the right were classified correctly only by the proposed PIDL and misclassified by the AE model.

5. Conclusions

Scarcity of condition monitoring data is a common challenge for practical deployment of fault detection algorithms. Data scarcity may be due to missing data, due to a low time resolution of the data or due to a low spatial resolution. The latter is a common situation in large scale PV power plants, in which condition monitoring data is often available at a low spatial granularity level, e.g. aggregating the monitored power production over a large number of individual assets. However, a similar situation applies to other large infrastructures, where the data volume is often reduced using a more coarse-grained aggregation when monitoring the assets.

In order to enable high fidelity fault detection despite the data scarcity challenge, we introduced a physics-informed artificial intelligence algorithm. With this approach, physical information is exploited in order to transfer the data augmentation from a domain with abundant data to a domain with scarce data. We demonstrated the high performance of the algorithm on operational data from a PV power plant with a low data granularity, and showed its clear superiority over a purely data-driven approach. Moreover, we showed that its performance is similar to our previous results achieved on a high data granularity power plant. Future research directions include an extension of the approach to additional fault and power loss mechanisms, aiming at effective diagnostics of the power loss root cause.

Acknowledgment

This research was funded by Innosuisse - Swiss Innovation Agency under grant No. 55018.1 IP-ICT.

References

Amaral, T. G., Pires, V. F., & Pires, A. J. (2021). Fault detection in pv tracking systems using an image processing algorithm based on pca. *Energies*, *14*(21), 7278.

Bansal, P., Zheng, Z., Shao, C., Li, J., Banu, M., Carlson, B. E., & Li, Y. (2022). Physics-informed machine learning assisted uncertainty quantification for the corrosion of dissimilar material joints. *Reliability Engineering & System Safety*, *227*, 108711.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2019). Hybrid deep fault detection and isolation: Combining deep neural networks and system performance models. *arXiv preprint arXiv:1908.01529*.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Chen, Z., Chen, Y., Wu, L., Cheng, S., & Lin, P. (2019). Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. *Energy Conversion and Management*, *198*, 111793.

Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., & Pavan, A. M. (2016). A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renewable Energy*, *90*, 501–512.

Daliento, S., Chouder, A., Guerriero, P., Pavan, A. M., Mellit, A., Moeini, R., & Tricoli, P. (2017). Monitoring, diagnosis, and power forecasting for photovoltaic fields: A review. *International Journal of Photoenergy*, *2017*.

Frank, S., Heaney, M., Jin, X., Robertson, J., Cheung, H., Elmore, R., & Henze, G. (2016). *Hybrid model-based and data-driven fault detection and diagnostics for commercial buildings* (Tech. Rep.). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Gao, W., & Wai, R.-J. (2020). A novel fault identification method for photovoltaic array via convolutional neural network and residual gated recurrent unit. *IEEE access*, *8*, 159493–159510.

Huber, L. G., Palmé, T., & Chao, M. A. (2023). Physics-informed machine learning for predictive maintenance: applied use-cases. In *2023 10th ieee swiss conference on data science (sds)* (pp. 66–72).

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Kohtz, S., Xu, Y., Zheng, Z., & Wang, P. (2022). Physics-informed machine learning model for battery state of health prognostics using partial charging segments. *Mechanical Systems and Signal Processing*, *172*, 109002.

Li, B., Delpha, C., Diallo, D., & Migan-Dubois, A. (2021). Application of artificial neural networks to photovoltaic fault detection and diagnosis: A review. *Renewable and Sustainable Energy Reviews*, *138*, 110512.

Li, W., Zhang, J., Ringbeck, F., Jöst, D., Zhang, L., Wei, Z., & Sauer, D. U. (2021). Physics-informed neural networks for electrode-level state estimation in lithium-ion batteries. *Journal of Power Sources*, *506*, 230034.

Mansouri, M., Trabelsi, M., Nounou, H., & Nounou, M. (2021). Deep learning based fault diagnosis of photovoltaic systems: A comprehensive review and enhancement prospects. *IEEE Access*.

Mellit, A., Tina, G. M., & Kalogirou, S. A. (2018). Fault detection and diagnosis methods for photovoltaic systems: A review. *Renewable and Sustainable Energy Reviews*, *91*, 1–17.

Pillai, D. S., & Rajasekar, N. (2018). A comprehensive review on protection challenges and fault diagnosis in pv systems. *Renewable and Sustainable Energy Reviews*, *91*, 18–40.

Racharla, S., & Rajan, K. (2017). Solar tracking system–a review. *International journal of sustainable engineering*, *10*(2), 72–81.

Rai, A., & Mitra, M. (2021). A hybrid physics-assisted machine-learning-based damage detection using lamb wave. *Sādhanā*, *46*(2), 64.

Rausch, R. T., Goebel, K. F., Eklund, N. H., & Brunell, B. J. (2005). Integrated In-Flight Fault Detection and Accommodation: A Model-Based Study. In *Volume 1: Turbo expo 2005* (pp. 561–569). ASME. doi: 10.1115/GT2005-68300

Triki-Lahiani, A., Abdelghani, A. B.-B., & Slama-Belkhodja, I. (2018). Fault detection and monitoring systems for photovoltaic installations: A review. *Renewable and Sustainable Energy Reviews*, *82*, 2680–2692.

Wu, Y., Sicard, B., & Gadsden, S. A. (2024). A review of physics-informed machine learning methods with applications to condition monitoring and anomaly detection. *arXiv preprint arXiv:2401.11860*.

Zgraggen, J., Guo, Y., Notaristefano, A., & Goren Huber, L. (2022). Physics informed deep learning for tracker fault detection in photovoltaic power plants. In *14th annual conference of the prognostics and health management society, nashville, usa, 1-4 november 2022* (Vol. 14).

Zgraggen, J., Guo, Y., Notaristefano, A., & Goren Huber, L. (2023). Fully unsupervised fault detection in solar power plants using physics-informed deep learning. In *33rd european safety and reliability conference (esrel), southampton, united kingdom, 3-7 september 2023* (pp. 1737–1745).

Data-Driven Prognostics with Multi-Layer Perceptron Particle Filter: a Cross-Industry Exploration

Francesco Cancelliere[1], Sylvain Girard [2], Jean-Marc Bourinet [3]

[1,2] *PHIMECA, Paris, 75012, France*
cancelliere@phimeca.com
girard@phimeca.com

[1,3] *SIGMA Clermont University, Aubiere, 63178 , France*
jean-marc.bourinet@sigma-clermont.fr

ABSTRACT

The integration of particle or Kalman filters with machine learning tools like support vector machines, Gaussian processes, or neural networks has seen extensive exploration in the context of prognostic and health management, particularly in model-based applications. This paper focuses on the Multi-Layer Perceptron Particle Filter (MLP-PF), a data-driven approach that harnesses the non-linearity of MLP to describe degradation trajectories without relying on a physical model. The Bayesian nature of the particle filter is utilized to update MLP parameters, providing flexibility to the method and accommodating unexpected changes in the degradation behavior.

To showcase the versatility of MLP-PF, this work demonstrates its seamless integration into diverse use cases, such as lithium-ion battery analysis, virtual health monitoring for turbofans, and the assessment of fatigue crack growth. We illustrate how it effortlessly accommodates various contexts through slight parameter modifications. Adjustment includes variation in the number of neurons or layers in the MLP, threshold adjustments, initial training refinements and the adaptation of the process noise. Addressing different degradation processes across these applications, MLP-PF proves its adaptability and utility in various contexts.

These findings highlight the method's versatility in adapting to diverse use cases and its potential as a robust prognostic tool across various industries. MLP-PF offers a practical and efficient means of estimating remaining useful life and predicting degradation in complex systems, with implications for advancing prognostic tools in diverse applications.

1. INTRODUCTION

Prognostic and Health Management (PHM) plays a crucial role in engineering by aiming to estimate the health state, detect early failures, and predict the remaining useful life of systems or components (Zio, 2022). Implementing PHM algorithms allows for condition-based or predictive maintenance strategies, ultimately optimizing maintenance frequency and reducing operational costs (Bailey, Sutharssan, Yin, & Stoyanov, 2015). Traditional physics-based methods in PHM rely on known equations, and demand extensive domain knowledge while generally being computationally expensive, limiting their real-time applicability (Chang, Fang, & Zhang, 2017).

In contrast, the rise of data availability in recent years coincides with the exploration of data-driven methods such as neural networks, random forests, and support vector machines (Wang, Jin, Deng, & Fernandez, 2021; Hu, Xu, Lin, & Pecht, 2020; Vanem et al., 2023). However, these methods often face challenges related to data quantity, quality, and generalization across unseen conditions. To overcome these hurdles, hybrid approaches have been proposed (Cancelliere, Girard, Bourinet, & Broggi, 2023; Li et al., 2024), aiming to combine the strengths of data-driven and physics-based methods.

Among the hybrid approaches, a common one consists of integrating particle or Kalman filters with machine learning tools like neural networks or support vector machines (Dong, Jin, Lou, & Wang, 2014; Jha, Bressel, Ould-Bouamama, & Dauphin-Tanguy, 2016). In these frameworks, machine learning tools act as surrogates for physics-based models, reducing the need for extensive domain expertise. Meanwhile, Bayesian filters allow to quantify the uncertainties associated with the prediction, enhancing the robustness of the approach. Other works, such as (Ma, Karkus, Hsu, & Lee, 2020) or (Ge, Sun, & Ma, 2019) proposed combination of PF with, respectively, recurrent neural network (RNN) and long-short term memory

network (LSTM). A more comprehensive review of combination of PF and data driven techniques can be found in (Reza et al., 2024).

A combination of radial basis functions with particle filters, initially proposed by (Sbarufatti, Corbetta, Giglio, & Cadini, 2018), presented a novel approach to estimate the state of charge of lithium-ion batteries, which was later extended to predict the end of life for batteries by replacing the surrogate model with a Multi-Layer Perceptron (MLP) neural network (Cadini, Sbarufatti, Cancelliere, & Giglio, 2019). This adaptation, called Multi-Layer Perceptron Particle Filter (MLP-PF), capitalizes on the non-linear nature of MLPs to describe system degradation trajectories and leverages the Bayesian framework of particle filters to adjust to incoming measurements.

The primary contribution of this work lies in the application of the MLP-PF to three different case studies. By demonstrating the versatility of MLP-PF, this study showcases its seamless integration into diverse applications, starting from the case of lithium-ion batteries, then changing to the estimation of a virtual health indicator for turbofans, and the assessment of fatigue crack growth. Through minor parameter modifications such as variations in MLP architecture, threshold adjustments, initial training refinements, and adaptation of process noise levels, MLP-PF effortlessly accommodates various contexts.

Differing from conventional data-driven approaches, this study adopts a single historical degradation trajectory as training for the MLP neural network. The training serves merely as a starting point for the (PF) to explore the state-space, relying on its Bayesian nature to discern the hidden degradation dynamics. This approach furnish the algorithm with exceptional adaptability while significantly mitigating the need for extensive historical data, a primary drawback of traditional data-driven methods.

To evaluate the algorithm's performance, various metrics including Relative Accuracy, Confidence Interval Coverage (Jules, Cancelliere, Mattrand, & Bourinet, 2023), and the β Metric (Lall, Lowe, & Goebel, 2013) are employed. These metrics assess not only accuracy and precision but also consider the uncertainty associated with the predictions, providing a comprehensive evaluation framework.

Addressing various degradation processes across different applications highlights the adaptability and utility of the MLP-PF. These findings emphasize the method's versatility in accommodating diverse use cases and underscore its potential as a robust prognostic tool across multiple industries. MLP-PF provides a practical and efficient means of estimating remaining useful life and predicting degradation in complex systems, thereby advancing prognostic tools across a broad spectrum of applications.

The structure of this paper is the following: Section 2 briefly describes the proposed method and the metrics used to evaluate performance. Following this, Section 3 introduces the three use cases addressed in this study, while Section 3.1, Section 3.2 and Section 3.3 present the results corresponding to each case. Finally, Section 4 draws conclusions and provides perspectives on this work.

2. Multi Layer Perceptron Particle Filter

The method employed in this work was first proposed by (Sbarufatti et al., 2018), where a combination of radial basis function neural networks and particle filters was used to estimate the state of charge of lithium-ion batteries. The method was later improved by (Cadini et al., 2019), where it was extended to estimate the state of health of the battery. In this work, the method is applied to three different use cases to showcase its ability to adapt to different contexts with slight changes in the hyperparameters.

The multi-layer perceptron neural network is used as a surrogate for the given degradation model, such as the turbofan VHI or the batteries' capacity. In all cases, it consists of a single input, which is the discrete time step k, and a single output $\tilde{g}$, representing the predicted value at the given time step. The decision to use a simple neural network, such as an MLP, is driven by the necessity for flexibility and the desire to minimize the number of parameters estimated by the PF. Despite its simplicity, an MLP remains capable of capturing the nonlinearities inherent in the data. This choice strikes a balance between model complexity and computational efficiency, enabling effective integration with the PF framework.

The internal architecture of the network (number of layers, number of neurons per layer, and the activation functions) is case-dependent, particularly on the shape of the degradation trajectories and to ensure computational times are compatible with the given context (higher the network complexity, higher the computational time). The parameters of the network, meaning its weights and biases, are then packed into a vector x_k. The starting parameters x_0 are obtained by training the network based on a known run-to-failure degradation process, as can be observed in the (a) figures of the three cases.

The particle filter (Arulampalam, Maskell, Gordon, & Clapp, 2002) is a sequential Monte-Carlo algorithm that generates a set of particles which are used to estimate the posterior probability density function (PDF) of a hidden state, which in our case are the parameters x_k. Hence, a set of N_s copies (i.e particles) of x_0 is generated based on:

$$x_k^i = x_{k-1}^i + \omega_{k-1} \tag{1}$$

where i is the index of the particles and ω is the process noise, which is an hyperparameter that has to be carefully tuned. Each x_k^i contains the parameters of a MLP, which,

when propagated through the network, generates a prediction of a possible degradation trajectory. The PF operates by applying a prediction-update recurrence. The predictions at time step $k-1$ serve as the prior PDF, which is updated at each subsequent time step k upon the arrival of new observations. The update of the particles is performed by computing their likelihood, which indicates how close the i^{th} degradation trajectory is to the actual observations. The likelihood $\mathcal{L}_k^i$ of each particle i is computed as:

$$\mathcal{L}_k^i = p(z_{0:k}|x_k^i) = ((2\pi)^{k+1}|\Sigma_\eta|)^{-0.5} \exp\left\{-\frac{1}{2}\left(z_{0:k} - \tilde{g}(x_k^i, 0:k)\right)^{\mathrm{T}} \Sigma_\eta^{-1} \left(z_{0:k} - \tilde{g}(x_k^i, 0:k)\right)\right\} \tag{2}$$

where $z_{0:k}$ are the observations from 0 to k, $\tilde{g}(x_k^i, 0:k)$ is the prediction of the network for the i^{th} particle and Σ_η is the diagonal covariance matrix, with diagonal element equal to η, representing the measurement noise and assumed Gaussian. $\mathcal{L}_k^i$ is the probability of obtaining the measurement z_k given the i^{th} prediction $\tilde{g}_k$, and it is used as importance weight w_k^i for the particles.

To finally construct the posterior pdf the sampling importance resampling (SIR) algorithm is employed (Doucet, Godsill, & Andrieu, 2000): the weights are normalized, and the particles are resampled based on the normalized weights $\tilde{w_k}$. The closer the prediction is to the measurements, the higher the normalized importance weight of that particle, meaning that the particle is more likely to be resampled, which signifies that it is closer to the actual degradation trajectory of the observed process.

The importance weights of the particles are also utilized to enforce specific conditions, ensuring that particles adhere to desired behaviors. One example is imposing the monotonicity of the trajectory or setting bounds on the output value (e.g., ensuring it is always greater than 0). If a particle violates the specified condition, its weight is set to zero, indicating that it will not be resampled. Instead, it is replaced by a particle with a higher importance weight, thereby maintaining compliance to the desired conditions.

The collection of normalized particles at time step k, each representing a potential degradation trajectory, enables the computation of the posterior probability density function of the degradation state in future time steps. Consequently, it becomes possible to calculate statistics related to predictions, such as the mean and relative uncertainties. Additionally, by establishing a threshold for the end of life of the system, it becomes feasible to determine the distribution of the End of Life $p(\mathrm{EOL}_k|z_{0:k})$, and consequently the RUL_k as:

$$\mathrm{RUL}_k = \mathrm{EOL}_k - k \tag{3}$$

To evaluate the performance of the algorithm we use three different metrics. The first one is the cumulative relative accuracy, defined as:

$$\mathrm{CRA} = \frac{1}{T_{fail}} \sum_{k=0}^{T_{fail}} \left(1 - \left|\frac{\mathrm{RUL}_k^{\text{actual}} - \mathrm{RUL}_k^{\text{pred}}}{\mathrm{RUL}_k^{\text{actual}}}\right|\right) \tag{4}$$

where T_{fail} is the time step at which the system fails. This represents the distance of the prediction to the actual EOL, evaluated at each time step. A perfect prediction has a value of 1.

The second metric is the confidence interval coverage (Jules, Cancelliere, et al., 2023), which is used to assess the prediction considering the confidence interval, and is defined as:

$$CIC = \frac{1}{T_{fail}} \sum_{k=0}^{T_{fail}} \mathbb{1}_{RUL_k^{actual} \in \widehat{CI_k}} \tag{5}$$

where $\mathbb{1}_{RUL_k^{actual} \in \widehat{CI_k}}$ is the indicator function that takes one if the actual RUL lies in the predicted confidence interval, 0 otherwise. If the prediction at each time step include the RUL^{actual}, the CIC is going to be 1, while 0 if the true RUL is always outside the confidence interval.

The last indicator is called the β metric (Lall et al., 2013), which represent the area of the predictions that falls inside the α bound.

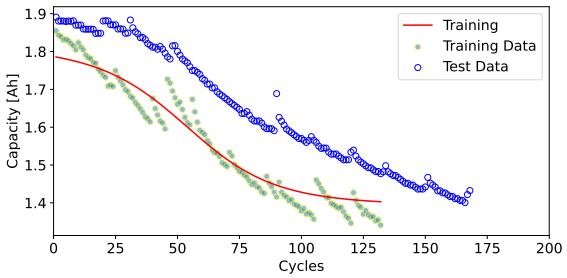
$$\beta_k = \frac{1}{T_{fail}} \sum_{k=0}^{T_{fail}} \int_{\mathrm{RUL}_k - \alpha}^{\mathrm{RUL}_k + \alpha} \mathrm{PDF}(\mathrm{RUL})\, d\mathrm{RUL} \tag{6}$$

The α bounds are defined as $RUL^{actual} \pm \alpha$. The β metric evaluates the accuracy of predicted RUL bounds compared to true RUL bounds, considering a specified uncertainty level (α). It quantifies the overlap between predicted and true RUL bounds normalized by the true RUL length. Higher values indicate better agreement between predicted and true bounds, reflecting improved prediction accuracy.

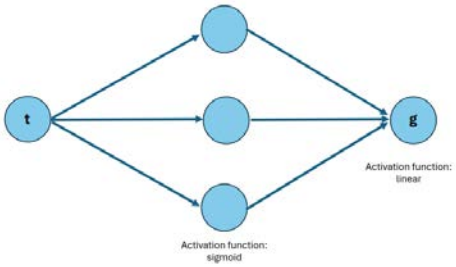
3. USE CASES

The proposed approach will now be applied to three use cases: estimating the end of life of lithium-ion batteries based on their decreasing capacity, propagating a virtual health indicator developed to estimate the state of health of turbofans, and modeling the growth of a fatigue crack in a panel. Although these cases share a time-dependency, their degradation processes differ significantly in terms of shape and rapidity. Therefore, we employ three different MLP network architectures, each tailored to the specific characteristics of its respective case.

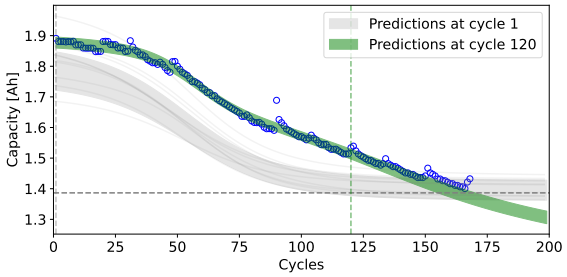
For consistency and comparison, in each cases we use the same number of particles, $N_s = 1000$, and the same number of epochs for the initial training (epochs = 500). Additionally, we employ a decreasing variance, defined as:



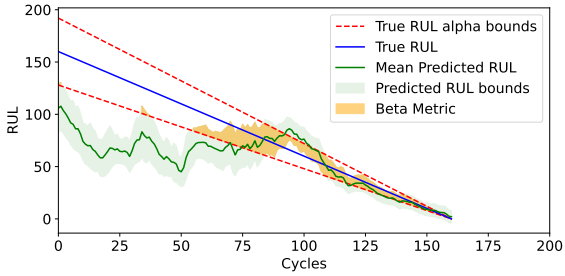
(a) Training and testing dataset for Li-Ion batteries. The red line is the output of the trained MLP.



(b) Architecture of the MLP neural network used for the Li-Ion batteries case.



(c) Two instant of time of the simulation. The grey dotted line is the end of life threshold.



(d) Remaining Useful Life prediction in terms of mean and uncertainty bounds.

Figure 1. Lithium-Ion Battery case.

$$\omega_k = \sigma_0 e^{-\frac{k}{\sigma_1}} + \sigma_2 \tag{7}$$

Here, ω perturbs the MLP parameters, as described in Eq. (1).

The use of a decreasing process variance is crucial for the convergence of the algorithm. Initially, a higher variance (which practically signifies a higher perturbation of x_k in Eq. (1)) is necessary to explore the state space and adapt to the first incoming observations, especially if these are significantly different from the training data. As more observations become available, the variance is reduced to reflect the increased information about the actual system. This reduction in variance helps prevent a single observation, especially a noisy one, from excessively perturbing the prediction. This strategy ensures a balanced adaptation process, enabling the algorithm to remain robust against noisy observations while gradually refining its predictions.

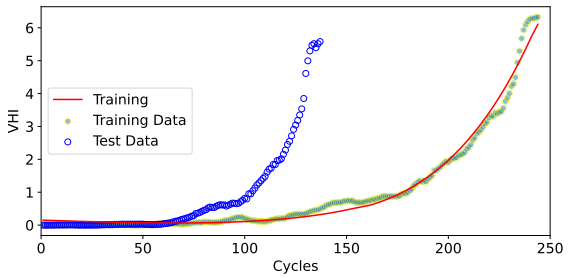
Given that the complexity of the network correlates with the number of parameters, using the same ω value for different architectures will lead to different perturbation. Specifically, higher complexity requires lower perturbation (i.e. lower ω) to prevent degeneration. If the MLP parameters change too rapidly, they may lose meaning and connection with prior information. Hence, the σ parameters and the measurement noise η, responsible for computing the particule likelihood in Eq. (2), vary across different cases.

Furthermore, in all three cases, we opt to use a single trajectory for the initial training. This choice aims to demonstrate the algorithm's ability to adapt to varying conditions and its capacity to achieve satisfactory performance in predicting the RUL without requiring a large amount of data. The proposed metrics are evaluated throughout all the degradation process and in the last 25% of life. This evaluation demonstrates that the algorithm's performance improves over time as more information becomes available, and it converges to the target data even when the initial training data differ significantly.

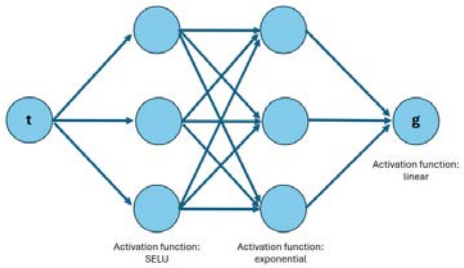
3.1. Lithium-ion Batteries

The dataset used for the first use case is the one developed by NASA for the prognostic and diagnostic analysis of batteries (Saha & Goebel, 2007). The capacity of batteries decreases over time due to usage and electrochemical reactions occurring inside the battery. The end of life of batteries is typically defined when the capacity drops below 80% of the initial capacity. However, to make the most of the dataset, in this work, we set a threshold of 1% higher than the last point, which is 1.42 Ah. In Fig. 1(a) the two batteries used for the initial training of the network (battery 18 of the dataset) and for testing (battery 7) are shown.

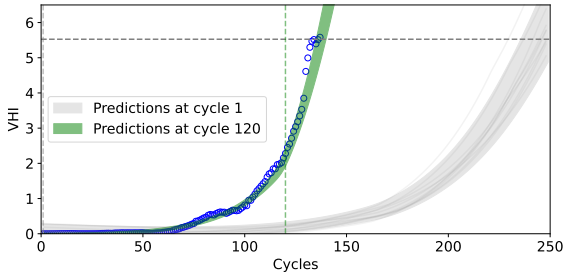
The architecture of the network consists of a single hidden layer with 3 neurons, where the activation functions are a sigmoid for the hidden layer and linear for the output layer. The network structure is depicted in Fig. 1(b). This results in a total of 6 weights and 4 biases, which, after training, are stacked



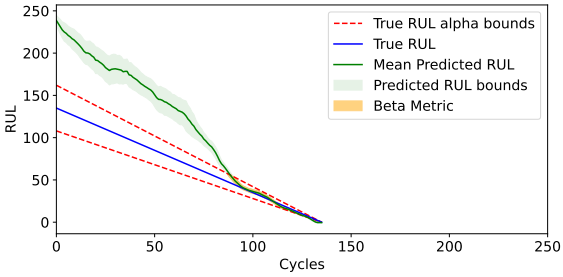
(a) Training and testing dataset for VHI. The red line is the output of the trained MLP.



(b) Architecture of the MLP neural network used for the VHI case.



(c) Two instant of time of the simulation. The grey dotted line is the end of life threshold.



(d) Remaining Useful Life prediction in terms of mean and uncertainty bounds.

Figure 2. Virtual Health Indicator case.

in the vector x_0, resulting in 10 parameters. The output of the trained network with these parameters is represented by the red line in Fig. 1(a).

The initial process noise is taken as $\sigma_0 = 5 \times 10^{-2}$, while the floor value is set to $\sigma_2 = 10^{-5}$, with a decreasing rate σ_1 of 50. The measurement noise is set to $10^{-}2$. Due to the simplicity of the network, the initial value σ_0 is relatively high, providing more flexibility to the algorithm. Furthermore, the adaptability of the algorithm is necessary due to the intrinsic nature of lithium-ion batteries, which can perform differently from one another, as observed in Fig. 1(a). The initial perturbation, obtained by applying Eq. (1) to each of the N_s particle can be observed in Fig. 1(c) as the grey lines.

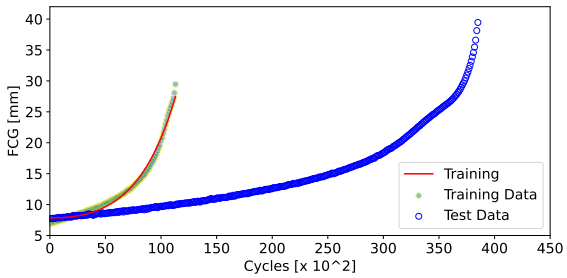
The results of the simulation are presented in Fig. 1(c) and Fig. 1(d). In the first, two instances of time, at the beginning and about the end of the simulation, are shown, highlighting the adaptability of the algorithm. Starting from the initial training, the algorithm adapts to incoming measurements and estimates the new degradation behavior. The last figure shows the results in terms of remaining useful life estimation. Initially, the predictions were more related to the training data, which has a faster end of life, while converging to the actual RUL at about the halfway point of the battery's lifetime. The evaluation of the algorithm's performance is reported in Table 1, where it can be observed that all the metrics improved when evaluated in the last 25% of the lifetime. Particularly, the Confidence Interval Coverage 25 has a value of 1, indicating that the actual RUL has always been inside the predicted bounds.

3.2. Virtual Health Indicator

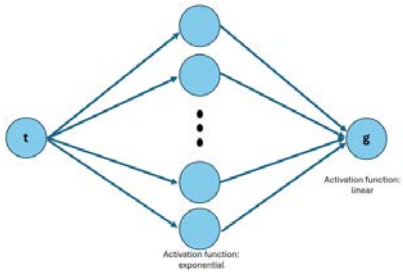
The second use case proposed here involves the estimation of the future behavior of a virtual health indicator developed for estimating the state of health of turbofans (Jules, Mattrand, & Bourinet, 2023). This VHI measures the degradation of turbofans, thus, opposite to the batteries case, it exhibits an upward trajectory, where a higher value indicates higher degradation. Similarly to the previous case, the end-of-life threshold has been set to utilize the maximum available number of cycles from the test dataset.

Upon observing the historical data of the VHI, it can be noted that initially, it exhibits a flat trajectory, remaining nearly at zero until the degradation process begins, after which it adopts an exponential-like trajectory. To accommodate this, the proposed network for this case consists of two hidden layers with 3 neurons each. The first layer employs a scaled exponential linear unit (SELU) activation function, while the second layer employs an exponential activation function. The output layer uses a linear activation function. The structure of this network (see Fig. 2(b)) is thus more complex, consisting of a total of 22 parameters (15 weights and 7 biases).

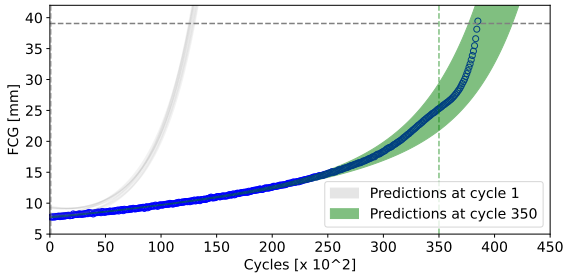
As mentioned in Section 2, certain conditions can be enforced on the particles to help them meet specific constraints. In this case, since the VHI has been designed to be greater than 0



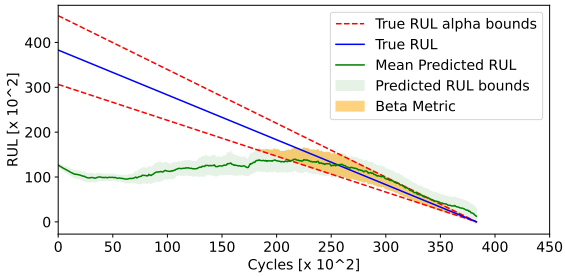
(a) Training and testing dataset for FCG. The red line is the output of the trained MLP.



(b) Architecture of the MLP neural network used for the FCG case with 20 neurons in the hidden layer.



(c) Two instant of time of the simulation. The grey dotted line is the end of life threshold.



(d) Remaining Useful Life prediction in terms of mean and uncertainty bounds.

Figure 3. Fatigue Crack Growth case.

and monotonous, these two conditions have been enforced by eliminating particles at each iteration that did not adhere to them. With more parameters, the initial process noise had to be slightly reduced, and particularly, we set the value of $\sigma_0 = 10^{-2}$. The other values (floor noise σ_2 and the measurement noise η) remained unchanged.

Fig. 2(a) displays the training and test data, as well as the output of the trained network. It can be observed that the shape of the two trajectories is similar; however, the rate of degradation varies notably, with the training dataset exhibiting a slower trend. Additionally, the initial flat plateau adds complexity to the prediction task since the algorithm receives measurements close to the expected values, resulting in high likelihood. This behavior is illustrated in Fig. 2(d), where initially, the predicted RUL decreases almost constantly, indicating little variation in prediction. Once the measurements from the VHI start to increase, signaling the onset of degradation, the algorithm quickly adapts to the new degradation behavior and converges to the actual RUL.

Even in this case, all metrics improved when evaluated in the last quarter of the lifetime. We note a relatively low CIC in this case, which can be attributed to the narrow prediction bounds. On the other hand, this led to a relatively high value of the β Metric of the last quarter.

3.3. Fatigue Crack Growth

The last use case concerns the propagation of a crack in a rectangular plate of commercial 316L steel (Langlois Raphael, 2018) subjected to a fatigue load. The dataset comprises two tests of identical plates, with a cycling tensile loading applied with a frequency of 10 Hz and a R ratio of 0.1. The first one is subjected to a maximum force of 15 kN (test data, see Fig. 3(a)) and the second to a maximum force of 22.5 kN (training data, also in Fig. 3(a)), with a As expected, the higher the applied force, the faster the crack propagates.

The trajectories follow an exponential-like function. Therefore, the proposed architecture for this problem consists of a single hidden layer with an exponential activation function (see Fig. 3(b)). To enforce the exponential behavior of the MLP (and also to challenge the algorithm), the hidden layer consists of 20 neurons, nearly tripling the number of parameters to 61.

Due to the higher number of parameters, the values of σ_0 and σ_2 have to be significantly decreased. For this simulation, they have been set to $\sigma_0 = 5 \times 10^{-4}$ and $\sigma_2 = 10^{-6}$. As in the VHI case, the monotonicity of the curve is enforced, particularly since this is a physical constraint.

In contrast to the VHI case, we use the faster degradation as training data while attempting to estimate the slower trajectory. This poses a challenge for the algorithm since extrapolating future data without prior examples is inherently

Table 1. RUL Evaluation Metrics.

	Li-Ion	VHI	FCG
CRA	0.703	0.474	0.489
CRA 25	0.772	0.762	0.240
CIC	0.475	0.140	0.425
CIC 25	1.000	0.294	1.000
β	0.326	0.277	0.273
β 25	0.465	0.737	0.404

complex for data-driven algorithms. Nonetheless, even in this case, we observe that the algorithm adapts quite rapidly, with the remaining useful life initially remaining constant (see Fig. 3(d)), indicating that the algorithm recognized early on that the training degradation was faster. As the correct RUL is approached around the halfway point of the lifetime, the algorithm is able to capture the new trajectory and remains consistent with the prediction.

In terms of metric, we note that the CRA is low, especially in the last 25%. This is mainly due to the relative error in the very last points, where even a small error in the average predicted RUL leads to a significant penalization of the CRA, as the actual RUL is a small number. In contrast, we observe a perfect coverage in the last quarter, as the actual RUL has always fallen within the predicted bounds during that period.

4. Conclusions

The proposed methodology, combining multi-layer perceptron neural networks and particle filters, demonstrated its adaptability and effectiveness in estimating the remaining useful life across diverse engineering systems. By applying the method to three distinct use cases – estimating the end of life of lithium-ion batteries, predicting the behavior of a Virtual Health Indicator in turbofans, and analyzing the propagation of fatigue cracks in steel plates – we showcased its versatility and accuracy in capturing degradation processes. The utilization of a single training history for each case underscores the robustness of the algorithm and its adaptability even when limited data about the target system are available. Evaluation metrics such as Cumulative Relative Accuracy (CRA), Confidence Interval Coverage (CIC), and the β metric provided valuable insights into the accuracy, coverage, and uncertainty of predicted RUL bounds. These findings emphasize the practical implications of accurate RUL estimation in predictive maintenance, enabling proactive decision-making to optimize maintenance schedules and reduce operational costs. Further research can explore enhancements to the methodology and its application to additional use cases beyond time-dependent applications, thus enhancing its utility and effectiveness in real-world scenarios

Acknowledgment

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 955393

References

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/nongaussian bayesian tracking. *IEEE Transaction on Signal Processing*, *50*(2), 174–188. doi: 10.1109/9780470544198.ch73

Bailey, C., Sutharssan, T., Yin, C., & Stoyanov, S. (2015). Prognostic and health management for engineering systems: a review of the data-driven approach and algorithms. *The Journal of Engineering*(July). doi: 10.1049/joe.2014.0303

Cadini, F., Sbarufatti, C., Cancelliere, F., & Giglio, M. (2019). State-of-life prognosis and diagnosis of lithium-ion batteries by data-driven particle filters. *Applied Energy*, *235*(June 2018), 661–672. doi: 10.1016/j.apenergy.2018.10.095

Cancelliere, F., Girard, S., Bourinet, J.-M., & Broggi, M. (2023). Grey-box Approach for the Prognostic and Health Management of Lithium-Ion Batteries. *Annual Conference of the PHM Society*, *15*(1), 1–8. doi: 10.36001/phmconf.2023.v15i1.3506

Chang, Y., Fang, H., & Zhang, Y. (2017). A new hybrid method for the prediction of the remaining useful life of a lithium-ion battery. *Applied Energy*, *206*, 1564–1578. doi: 10.1016/j.apenergy.2017.09.106

Dong, H., Jin, X., Lou, Y., & Wang, C. (2014). Lithium-ion battery state of health monitoring and remaining useful life prediction based on support vector regression-particle filter. *Journal of Power Sources*, *271*, 114–123. doi: 10.1016/j.jpowsour.2014.07.176

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*, 197–208. doi: 10.1023/A:1008935410038

Ge, Y., Sun, L., & Ma, J. (2019). An Improved PF Remaining Useful Life Prediction Method Based on Quantum Genetics and LSTM. *IEEE Access*, *7*, 160241–160247. doi: 10.1109/ACCESS.2019.2951197

Hu, X., Xu, L., Lin, X., & Pecht, M. (2020). Battery Lifetime Prognostics. *Joule*, *4*(2), 310–346. doi: 10.1016/j.joule.2019.11.018

Jha, M. S., Bressel, M., Ould-Bouamama, B., & Dauphin-Tanguy, G. (2016). Particle filter based hybrid prognostics of proton exchange membrane fuel cell in bond graph framework. *Computers and Chemical Engineering*, *95*, 216–230. doi:

10.1016/j.compchemeng.2016.08.018

Jules, E., Cancelliere, F., Mattrand, C., & Bourinet, J.-M. (2023). Remaining useful life prediction of turbofans with virtual health indicator: A comparison of particle filter-based approaches. , 75-82. doi: 10.1109/ICSRS59833.2023.10381439

Jules, E., Mattrand, C., & Bourinet, J.-M. (2023). Similarity learning for predictive maintenance: health indicator construction based on siamese neural networks and contrastive loss. *[Under Review]*.

Lall, P., Lowe, R., & Goebel, K. (2013). Prognostic health monitoring for a micro-coil spring interconnect subjected to drop impacts. *PHM 2013 - 2013 IEEE International Conference on Prognostics and Health Management, Conference Proceedings*(June 2013). doi: 10.1109/ICPHM.2013.6621458

Langlois Raphael, R. J., Coret Michel. (2018, November). Fatigue Crack Propagation Benchmark, GDR 3651 FATACRACK. Retrieved from `https://doi.org/10.5281/zenodo.1478472` doi: 10.5281/zenodo.1478472

Li, T., Chen, J., Yuan, S., Zarouchas, D., Sbarufatti, C., & Cadini, F. (2024). Particle filter-based fatigue damage prognosis by fusing multiple degradation models. *Structural Health Monitoring*, *0*(0), 14759217231216697. doi: 10.1177/14759217231216697

Ma, X., Karkus, P., Hsu, D., & Lee, W. S. (2020). Particle filter recurrent neural networks. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 5101–5108. doi: 10.1609/aaai.v34i04.5952

Reza, M. S., Mannan, M., Mansor, M., Ker, P. J., Mahlia, T. M., & Hannan, M. A. (2024). Recent advancement of remaining useful life prediction of lithium-ion battery in electric vehicle applications: A review of modelling mechanisms, network configurations, factors, and outstanding issues. *Energy Reports*, *11*(April), 4824–4848. doi: 10.1016/j.egyr.2024.04.039

Saha, B., & Goebel, K. (2007). *Battery Data Set. NASA Ames Prognostics Data Repository.*

Sbarufatti, C., Corbetta, M., Giglio, M., & Cadini, F. (2018). Adaptive prognosis of lithium-ion batteries based on the combination of particle filters and radial basis function neural networks. *Journal of Power Sources*, *344*, 128–140. doi: 10.1016/j.jpowsour.2017.01.105

Vanem, E., Liang, Q., Ferreira, C., Agrell, C., Karandikar, N., Wang, S., … Kandepu, R. (2023). Data-Driven Approaches to Diagnostics and State of Health Monitoring of Maritime Battery Systems. *Annual Conference of the PHM Society*, *15*(1), 1–17. doi: 10.36001/phmconf.2023.v15i1.3437

Wang, S., Jin, S., Deng, D., & Fernandez, C. (2021). A Critical Review of Online Battery Remaining Useful Lifetime Prediction Methods. *Frontiers in Mechanical Engineering*, *7*(August), 1–19. doi: 10.3389/fmech.2021.719718

Zio, E. (2022). Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering and System Safety*, *218*(PA), 108119. doi: 10.1016/j.ress.2021.108119

Data-Driven Remaining Useful Life Estimation Approach for Neutron Generators in Multifunction Logging-While-Drilling Service

Karolina Sobczak-Oramus[1], Ahmed Mosallam[2], Nannan Shen[3], and Fares Ben Youssef [4]

[1] *SLB, Nowogrodzka Street 68, Warsaw, Mazowieckie, 02-014 Poland*
KSobczak@slb.com

[2,3] *SLB, 1 Rue Henri Becquerel, 92140 Clamart, France*
AMosallam@slb.com
NShen@slb.com

[4] *SLB, 135 Rousseau Road, 70592 Youngsville, Louisiana, United States*
FYoussef@slb.com

ABSTRACT

This paper introduces a data-driven approach for estimating the remaining useful life of the neutron generator component in logging-while-drilling tools. The approach builds on identification of the incipient failure modes of the neutron generator and constructing a health indicator that serves as a statistical representation of the component's deterioration over time. Afterwards, a K-nearest neighbors algorithm is trained to establish the relationship between the extracted health indicator values and the corresponding remaining useful life. The effectiveness of the presented approach is verified through the utilization of real-world data gathered from oil well drilling operations. The study is part of a long term project aimed at developing a digital fleet management system for drilling tools.

1. INTRODUCTION

The multifunction logging-while-drilling (LWD) tool shown in Fig. 1 is an industry-leading formation evaluation technology developed for oil well drilling applications (SLB, 2023).

Figure 1. Multifunction LWD service

During the drilling operations, this LWD tool collects information related to formation evaluation as well as equipment diagnostics, and some of this information is transmitted in real time via mud pulses, depending on the operational requirements. The full quantity of raw information is stored in a memory board for additional analysis upon the tool returning to surface after completing drilling operations.

This multifunction LWD tool integrates functionalities that were previously achieved by two or three LWD tools, resulting in significantly reduced drilling rig operating time, fewer electrical and communication failures due to physical tool-to-tool connections, and improved geological data quality from simultaneous and co-located measurements.

One of the other critical factors that places the technology ahead of all other competitors is the inclusion of a pulsed neutron generator (PNG), as shown in Fig. 2. The PNG is an electron neutron generator that produces high-energy neutrons five times higher in energy than conventional americium-beryllium chemical sources. The PNG produces neutrons by particle fusion reaction. Multiple security locking logic has been added to the PNG firmware and hardware and is contained in operation manuals, making the use and transportation of the PNG safe. The high-energy neutron emission also allows for a variety of additional and advanced formation measurements for the customer.

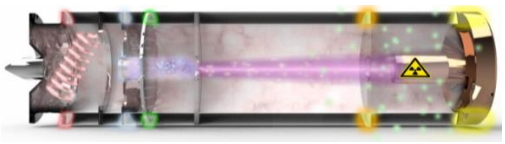


Figure 2. Pulsed neutron generator

The electrical and physical complexity of the PNG system makes it very demanding for a technician to become proficient in maintaining and troubleshooting the system and a failure of analysis would potentially result in critical field operation failures and jeopardize the company's reputation. Therefore, developing an automated fault diagnosis and degradation assessment tool to determine the health status of PNGs accurately and consistently is essential. This assessment tool can significantly reduce the potential for human error and enable users to make efficient and effective decisions (Zhan, Ahmad, Heuermann-Kuehn, & Baumann, 2010) (Isermann, 2006). This paper will focus on building data-driven remaining useful life estimation (RUL) for PNG systems, so appropriate actions can be taken for preventative replacement of the asset. The novelty and academic contributions of this study are highlighted by the fact that there has been no prior work related to RUL prediction for PNG systems, making this research a pioneering effort in the field.

The next section of this paper provides a detailed overview of the PNG system and discusses previous work and research relevant to this study. Research problem is formulated in the following section. The next two sections present the modeling approach and experimental results, respectively. A conclusions section completes the paper.

2. Pulsed Neutron Generator System

2.1. Description

For many years, the oil and gas industry has employed high-energy neutron generators in neutron-gamma-ray or neutron-neutron logging (Tittle, 1961). These generators offer several advantages over conventional chemical sources, including the ability to deactivate the PNG and eliminate radiation risks when not in use downhole. Additionally, the generators enable precise control over neutron output, facilitating more accurate measurements of formation properties.

In the field of nuclear well logging, achieving accurate formation measurements hinges on emitting neutron pulses to irradiate the Earth's formations and detecting the resulting radiation from the interaction between the Earth's formation atoms and the emitted neutrons. Understanding the characteristics of the neutron pulse, including its output and timing, is crucial for achieving precision. Ideally, the neutron pulse should exhibit a substantially square wave shape. The PNG, depicted in Fig. 3, proves instrumental in overcoming these technical challenges and facilitating the generation of desirable neutron pulses. Serving as a stand-alone particle accelerator, the PNG utilizes fusion reactions to produce neutrons.

Conducting a failure investigation and root cause analysis revealed two major failure modes of the PNG:

1. Internal cathode wire discontinuity due to overheating
2. Reduced neutron generation flux due to doped target wear

These failure modes can potentially compromise the functionality of the PNG and even lead to the failure of the LWD tool.

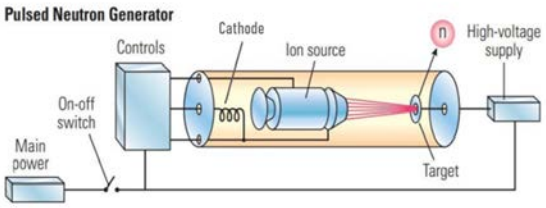


Figure 3. PNG architecture

2.2. Previous Work and Research

In prior research, a data-driven fault detection model for the PNG subsystem was introduced (Mosallam, Laval, Youssef, Fulton, & Viassolo, 2018). This approach involved creating a univariate representation known as a health indicator (HI). Subsequently, a classifier utilizing the decision tree method was trained to distinguish healthy and failed runs of PNGs. The method demonstrated high accuracy, providing quick and precise assessment for maintenance and field engineers.

Following that, a data-driven fault diagnostics method for the PNG system was introduced, specifically focusing on the detection of failures associated with the power supply boards (Mosallam, Kang, Youssef, Laval, & Fulton, 2023). This work complements the previously published fault detection model for the PNG subsystem (Mosallam et al., 2018) by providing detailed information indicating which electronic board or boards failed. The method extracts features from data channels capturing fault symptoms and builds support vector classifier models for each board. Experimental results showed an average accuracy of about 99% for all boards, reducing troubleshooting time and enabling automatic triggering of maintenance procedures for faulty boards.

The latest publication concentrated on the data-driven degradation modeling of the PNG system, emphasizing one of its incipient failure modes (Mosallam, Youssef, et al., 2023). The method extracts HI values from data channels quantifying component health degradation utilizing a random forest classification model. Experimental results demonstrate an average accuracy of 90.4%. The algorithm enables the identification of degradation stage of the PNG, empowering better planning for the equipment usage and avoidance of the failure during downhole operations.

However, in order to foster decision-making even more precisely, the necessity for a RUL estimation persists. This paper focuses on predicting the reduction of neutron generation flux due to doped target wear over time and determining the remaining useful time of the system. Integrating RUL estimation will enable proactive maintenance planning, better

decision-making on well sites, and future manufacturing forecasts and equipment delivery optimization based on worldwide RUL of active PNGs.

3. Problem Formulation

The main objective of prognostics is to minimize the equipment or system downtime by forecasting the RUL of the system (or critical components of the system), as shown in Fig. 4. The RUL prediction methods can be broadly classified into three categories: physics model based, data-driven, and hybrid (Medjaher, Tobon-Mejia, & Zerhouni, 2012)(Lei et al., 2018). Physics model based methods use mathematical models to describe the system's or component's physical behavior and predict its RUL. Although these methods require a deep understanding of the failure mechanisms and effective estimation of model parameters, they can provide accurate RUL estimation. On the other hand, data-driven methods use pattern recognition algorithms to learn patterns from historical data and make RUL predictions. The data-driven methods do not require a comprehensive understanding of the system failures but require high-quality data. Hybrid methods combine the strengths of both methods to improve RUL predictions.

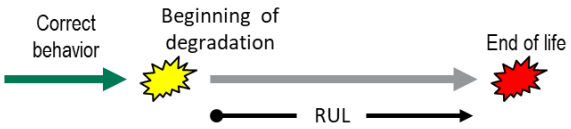


Figure 4. RUL forecast schematic

The PNG system being studied is highly complex, which limits the use of physics model-based and hybrid methods for predicting the PNG's (or PNG component's) RUL. Thus, the goal is to create a data-driven prognostic model that incorporates data related to the incipient failure modes of the PNG target. This model will estimate the target's RUL along with its confidence level as shown in Fig. 5.

There are two primary methods for building data-driven prognostic models: direct RUL mapping and cumulative degradation prognostics (Mosallam, Medjaher, & Zerhouni, 2016). The direct RUL mapping approach uses empirical models to directly correlate sensor data with the end of life (EOL) value, eliminating the need to determine the health status of the monitored component (see Fig. 6).

In contrast, the cumulative degradation prognostics approach uses empirical models to describe the system's degradation progression. This degradation information can then be used to estimate the health status of the system and predict the RUL based on the system's expected future behavior (see Fig. 7).

This study introduces a novel approach for predicting RUL using the direct RUL mapping approach. The primary objective is to establish a model that effectively captures the correlation between sensor measurements and end-of-life states, thereby enabling RUL prediction without relying on predefined alarm thresholds. Predefined alarm thresholds do not effectively capture the correlation between sensor measurements and EOL states, which can lead to less accurate predictions of RUL. The proposed method builds on the extraction of health indicators from historical training data, which serve as foundational reference models. Upon encountering new data, the approach employs a K-nearest neighbors (KNN) classifier to identify the most closely resembling HI within the database, subsequently leveraging it as a reliable RUL predictor.

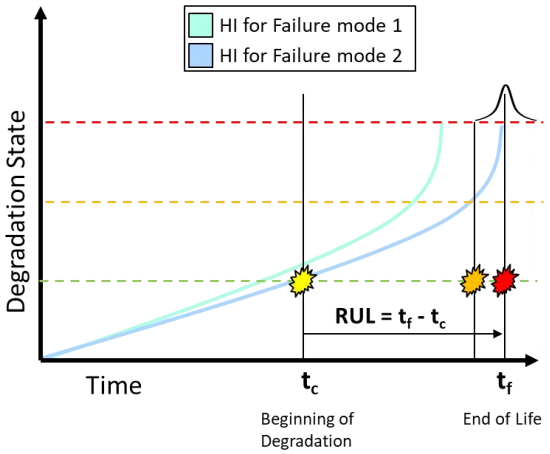


Figure 5. HIs for a system with two different failure modes

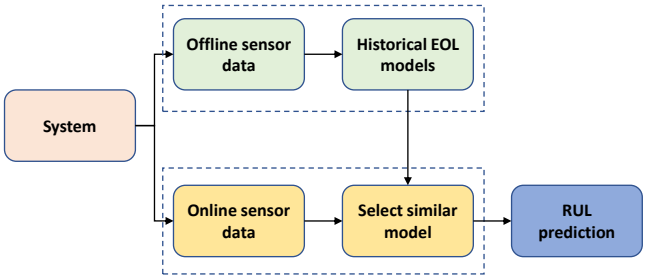


Figure 6. Direct RUL mapping approach

4. Proposed Method

The objective of the proposed method is to construct an HI based on sensor data that efficiently captures the PNG deterioration information. The HI values with RUL assignment are then modeled using a machine learning algorithm, which can estimate the RUL of the PNG. The proposed method is divided into four main steps: channel selection, preprocessing, HI construction, and modeling.

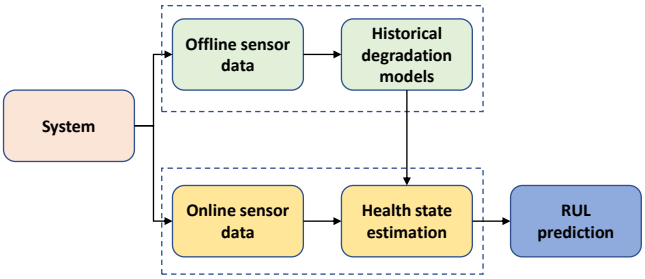


Figure 7. Cumulative degradation approach

4.1. Channel Selection

As highlighted in Section 1, the LWD tool generates a substantial number of high-resolution data channels during each drilling operation, leading to millions of data points. However, not all of these channels contribute information pertaining to the degradation of PNG over time. Enhancing the efficiency and precision of the HI involves the removal of irrelevant data channels. The selection of pertinent data channels relies on the expertise of subject matter experts with domain knowledge in nuclear physics and instrumentation. This process is crucial for optimizing the relevance of the data considered for the HI.

For the target failure mode, the following two data channels were selected:

- BLD: The internal high voltage in the PNG's minitron that enables the particles' acceleration to the target. The higher the BLD, the better the fusion reaction.
- BEAM: The particle beam current. The higher the BEAM, the higher the neutron emission.

Raw data for BLD and BEAM channels are presented in Fig. 8. The HI will be constructed using the data from the channels after the preprocessing step that will be described in the upcoming subsection. Note that the duration of each run is different according to the drilling job requirements, and the data of the sixth and eighth run before EOL are missing.

4.2. Preprocessing

The LWD data acquisition system commences recording data as soon as a field engineer initializes it for the forthcoming drilling operation (or run). The LWD tool follows the subsequent steps:

1. Tool initialization: the field engineer configures the acquisition parameters for the upcoming job, formats the tool memory, and begins the tool recording.
2. Shallow hole test: the field engineer confirms that the tool is functioning as expected inside the well before deploying the tool to the full well depth.

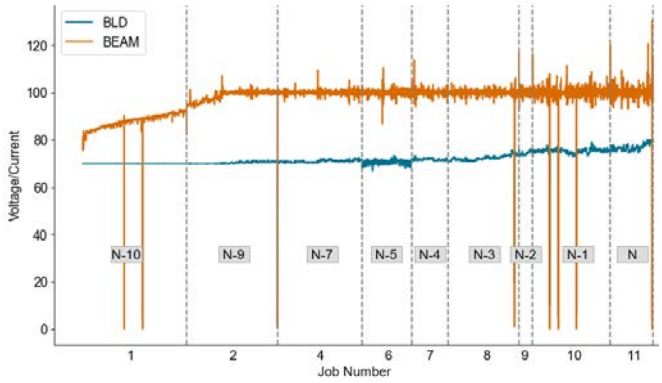


Figure 8. Raw data of the selected channels of eleven consecutive runs before EOL, where N denotes the last run, N-1 denotes the first run before EOL, N-2 denotes the second run before EOL, and so on.

3. Casing logging for caliper calibration: The field engineer calibrates the tool's ultrasonic measurement by using the known internal diameter of the metal casing connecting the rig to the wellbore and the known drilling fluid properties.
4. Drilling operation: The field engineer places the tool behind the drill bit for measurement acquisition during the physical drilling of the well.

For each run, the data collected during the initial three steps lack information pertaining to PNG degradation and are consequently excluded. Additionally, during periods when the PNG does not fire, the firmware generates dummy records to fill data channel gaps, known as missing values. All missing values must be disregarded as they do not provide any insights into faults. The duration of each drilling job varies. Let T represent the duration of a given drilling job. Consequently, the next step is undertaken to facilitate the extraction of a consistent statistical representation of the signal. Each time series of BLD and BEAM, denoted as

$$X = [x_1, \ldots, x_T]$$

and

$$Y = [y_1, \ldots, y_T]$$

respectively, generated throughout the run is segmented into $N = 200$ windows, determining the window size based on the duration of the run

$$w = \left\lfloor \frac{T}{N-1} \right\rfloor,$$

where each of $1, \ldots, N-1$ windows are of window size w and the last window is of a size $T \mod N-1$. Within each of 200 windows the minimal mode of each channel is extracted. To define the process of extracting the minimal mode, the function $g(Z)$ is denoted as the minimum of the mode of

Z, and is expressed as:

$$g(Z) = \min(\text{mode}(Z)).$$

For BLD channel, the minimal mode is extracted as follows:

$$x'_{i+1} = \begin{cases} g([x_{iw+1}, \ldots, x_{(i+1)w}]), & \text{if } i \in \{0, \ldots, N-2\} \\ g([x_{iw+1}, \ldots, x_T]), & \text{if } i = N-1 \end{cases}$$

resulting in

$$X' = [x'_1, \ldots, x'_N].$$

The same process is applied to BEAM channel, with the formula given by:

$$y'_{i+1} = \begin{cases} g([y_{iw+1}, \ldots, y_{(i+1)w}]), & \text{if } i \in \{0, \ldots, N-2\} \\ g([y_{iw+1}, \ldots, y_T]), & \text{if } i = N-1 \end{cases}$$

resulting in

$$Y' = [y'_1, \ldots, y'_N].$$

Finally, a median filter is applied to the sequence X' and Y' to smooth the signals, with $\tilde{X}$ and $\tilde{Y}$ denoting the smoothed X' and Y', respectively. Fig. 9 presents the raw signals of BLD and BEAM shown in Fig. 8 after preprocessing.

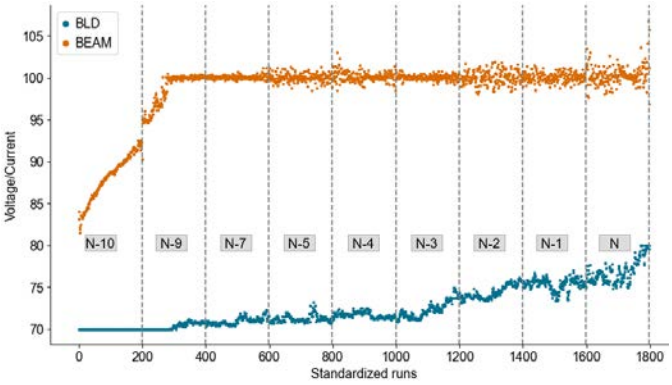


Figure 9. Preprocessed BLD and BEAM signals

4.3. Health Indicator Construction

In this algorithm step, an HI is constructed using the preprocessed data. The main objective of deriving this HI is to represent the system's degradation in a 1D array format. The HI is extracted as a result of element-wise addition of the preprocessed BLD and BEAM signals, which can be represented by the formula

$$HI = \tilde{X} + \tilde{Y}.$$

The decision to extract the HI by summing two channels comes from the fact that this approach ensures the HI has several key characteristics necessary for effectively monitoring health status of the PNG. These include sensitivity to degradation, monotonicity, predictive power, noise robustness, and consistency across conditions. Additionally, this method is computationally efficient and easily interpretable. Fig. 10 presents the HI values constructed using the Fig. 9 example data.

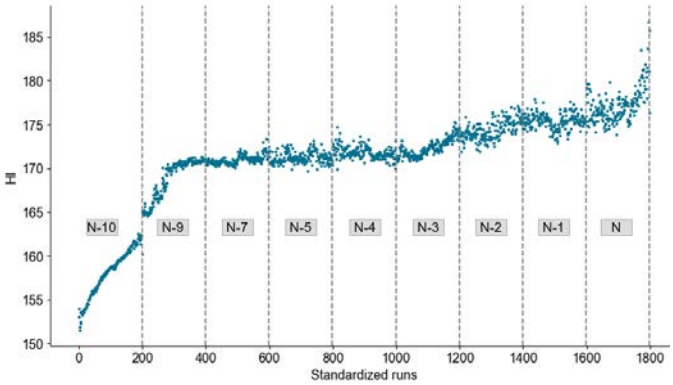


Figure 10. Constructed HI

It is important to highlight that the input data utilized reflects the incipient failure mode. Consequently, the resulting HI acts as a numerical gauge of the system's condition in a monotonic manner. Put simply, the HI offers a numerical representation of the system's health, enabling further analysis or decision-making processes.

4.4. Modeling

The HI constructed from the data collected during each drilling job results in 1D array representation, independent of the actual run duration. Thus, the RUL estimation problem can be converted into a regression problem if the array of each run has RUL assigned. To accomplish that, the usage of the PNG in hours estimated for each run is used, and respectively summed up over a lifespan. The formula for the RUL assignment is the following:

$$RUL_t = EOL - t, \forall EOL > t, \quad (1)$$

where t is the current time. The RUL labels serve as a target variable for each run. Specifically, this paper uses a KNN classifier to establish the relationship between the input HI values and their corresponding RUL i.e., $RUL_t = f(X_t)$, where

$$X_t = [x_{1,t}, x_{2,t}, \ldots x_{200,t}] \quad (2)$$

5. Experimental Results

A dataset containing operational data from 89 different LWD tool runs in different locations was collected to validate the proposed method. The dataset consists of historical runs from 16 different PNGs that reached the EOL. However, each PNG can operate different number of runs and the data availability highly varies between the PNGs. As mentioned in Section 2.2, in the previous work, the health state estimation model was developed (Mosallam, Youssef, et al., 2023). This algo-

rithm is applied for the PNG and enables distinguishing five degradation states: healthy, lightly degraded, moderately degraded, severely degraded and EOL. The dataset utilized for estimating the RUL differs from that used in previous studies, primarily due to an increased number of EOL cases captured in the current research. Previously, without a health state estimation model, the usage of the PNG until late stages of degradation was scarcely feasible, rendering RUL estimation impossible at that time. It is important to note that the main focus of this work is to predict the RUL when the PNG's degradation has already started, ad therefore, we train the model only from a certain degradation state. The RUL estimation is thus performed for the four stages of the PNG degradation being lightly degraded, moderately degraded, severely degraded or reaching EOL.

To assess the performance of the proposed method, the mean absolute percentage error (MAPE) evaluation metric is calculated. Let RUL_i be the actual remaining useful life and RUL_i^* be the predicted remaining useful life for $i = 1, \ldots, n$ where n is the total number of runs performed by all PNGs. The MAPE is defined as follows:

$$MAPE(\%) = \frac{100\%}{n} \cdot \sum_{i=1}^{n} \left| \frac{RUL_i - RUL_i^*}{RUL_i} \right| \quad (3)$$

To assess and compare the performance of the models, as well as to validate their ability to generalize the learned patterns, the dataset was split into a training set of 14 PNGs corresponding to 75 runs, and a test set of 2 PNGs resulting in 14 runs. Additionally, a specialized form of k-fold cross-validation is employed in the training phase. This method involves iterating through the PNGs, with each iteration leaving out one PNG for validation along with all its corresponding runs while training the model on the runs from the remaining $k - 1$ PNGs. This approach allows for efficient validation of how the model performs across various stages of PNG degradation, reducing the bias in performance estimation.

To select the best-performing model, two groups of algorithms were tested to estimate the RUL of the PNG: regressors and classifiers (Mosallam, 2014). The following algorithms were evaluated within each group: KNN, decision tree, random forest, and gradient boosting. Additionally, hyperparameter tuning was conducted to determine the best-performing algorithm.

The two best-performing algorithms: KNN and gradient boosting regressor (GBR) were selected for the further evaluation based on the overall MAPE presented in the Table 1. The selection of the single, most-suitable and best-performing algorithm is done based on both the MAPE overall and the MAPE calculated across various RUL intervals presented in Table 2. This approach ensures that the chosen model accurately predicts system behavior not only throughout its operational lifespan but also specifically as it approaches EOL conditions.

Table 1. MAPE of the LOOCV set for the trained algorithms

Algorithm name	MAPE (%)
K-Neighbors Classifier (K=2)	16.64 %
Gradient Boosting Regressor	17.26 %
K-Neighbors Classifier (K=3)	17.57 %
Random Forest Regressor	17.67 %
K-Neighbors Regressor (K=3)	17.78 %
K-Neighbors Regressor (K=2)	17.91 %
K-Neighbors Regressor (K=4)	18.40 %
K-Neighbors Classifier (K=4)	18.69 %
Random Forest Classifier	21.54 %
Decision Tree Classifier	21.92 %
Decision Tree Regressor	25.82 %
Gradient Boosting Classifier	30.18 %

Table 2. MAPE of the RUL intervals for the best performing algorithms

RUL interval	GBR MAPE (%)	KNN MAPE (%)
(0, 100]	26 %	13 %
(100, 200]	20 %	19 %
(200, 300]	16 %	20 %
(300, 400]	21 %	22 %
(400, 500]	8 %	11 %
(500, 600]	13 %	10 %
(600, 700]	6 %	5 %
(700, 800]	6 %	13 %

Taking into consideration the results of overall MAPE and MAPE calculated for the specified intervals, the KNN model outperforms other algorithms. The accuracy of prediction for the RUL in the interval $(0, 100]$ is significantly higher for KNN with $k = 2$, which is crucial as the PNG approaches the EOL. As previously stated, the objective is not solely to reduce the overall MAPE, but also to ensure the accuracy of predictions as the PNG approaches the end of its lifespan. The final phase of model validation involves training the model on the entire training set and evaluating its performance on the left-out test set. The MAPE for the test set is showcased in Table 3, confirming the consistent performance of both models.

Table 3. MAPE of the test set for the best-performing algorithms

Algorithm name	MAPE (%)
K-Neighbors Classifier (K=2)	16.24 %
Gradient Boosting Regressor	11.85 %

In conclusion, the KNN algorithm outperformed the GBR and other algorithms, as evidenced by the LOOCV outcomes and its superior accuracy in predicting the RUL of PNGs as they

approach the end of their lifespans. Furthermore, the KNN algorithm demonstrated strong performance on the test set, and offers greater explainability than GBR due to its straightforward methodology and direct reliance on training data points. Therefore, this model got selected for the implementation. Fig. 11 presents the RUL estimated by KNN model for all the available runs for one of the PNGs, resulting in MAPE of 11.1%.

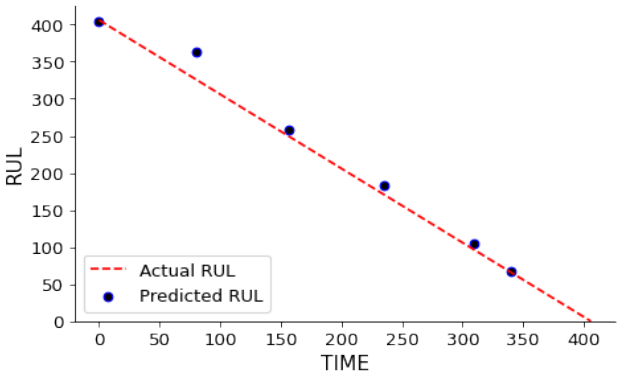


Figure 11. KNN prediction for selected PNG

6. Conclusions

This paper has presented a data-driven approach for RUL estimation of the PNG system in LWD tools. The method provides a quantitative measure of the component's deterioration by extracting the HI from BLD and BEAM data channels related to identified incipient failure mode of PNG. These HI values are used to build a KNN classification model to estimate the PNG RUL, which has been deployed as part of the health analyzer software for the LWD tool. Experimental results on actual operational data collected from the field resulted in the MAPE for LOOCV of 16.6%, and MAPE for the test set of 16.2%, demonstrating the effectiveness of the method. This method can assist maintenance engineers in promptly determining the PNG's remaining operational lifespan, reducing troubleshooting time; enable automatic triggers for maintenance activities for replacing faulty nuclear components; and improve decision making. Regarding future endeavors, there are intentions to address the second incipient failure mode of the PNG, with the aim of improving the prognostics model for the PNG.

References

Isermann, R. (2006). *Fault-diagnosis systems: An introduction from fault detection to fault tolerance*. Heidelberg: Springer-Verlag.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. doi: 10.1016/j.ymssp.2017.11.016

Medjaher, K., Tobon-Mejia, D. A., & Zerhouni, N. (2012, June). Remaining useful life estimation of critical components with application to bearings. *IEEE Transactions on Reliability*, *61*(2), 292–302. doi: 10.1109/TR.2012.2194175

Mosallam, A. (2014). *Remaining useful life estimation of critical components based on bayesian approaches.* (PhD dissertation). Université de Franche-Comté.

Mosallam, A., Kang, J., Youssef, F. B., Laval, L., & Fulton, J. (2023). Data-driven fault diagnostics for neutron generator systems in multifunction logging-while-drilling service. In *2023 prognostics and health management conference.*

Mosallam, A., Laval, L., Youssef, F. B., Fulton, J., & Viassolo, D. (2018). Data-driven fault detection for neutron generator subsystem in multifunction logging-while-drilling service. In *PHM society european conference.*

Mosallam, A., Medjaher, K., & Zerhouni, N. (2016). Data-driven prognostic method based on bayesian approaches for direct remaining useful life prediction. *Journal of Intelligent Manufacturing*, *27*, 1037–1048. doi: 10.1007/s10845-014-0933-4

Mosallam, A., Youssef, F. B., Sobczak-Oramus, K., Kang, J., Gupta, V., Shen, N., & Laval, L. (2023). Data-driven degradation modeling approach for neutron generators in multifunction logging-while-drilling service. In *2023 prognostics and health management conference.*

SLB. (2023, March). *EcoScope Multifunction LWD Service.* Retrieved from https://www.slb.com/drilling/surface-and-downhole-logging/logging-while-drilling-services/ecoscope-multifunction-lwd-service

Tittle, C. W. (1961). Theory of neutron logging I. *Geophysics*, *26*(1), 27-39. doi: 10.1190/1.1438839

Zhan, S., Ahmad, I., Heuermann-Kuehn, L., & Baumann, J. (2010, 09). Integrated PoF and CBM strategies for improving electronics reliability performance of downhole MWD and LWD tools. In *Spe annual technical conference and exhibition.* doi: 10.2118/132665-MS

Biographies

Karolina Sobczak-Oramus is a Senior Data Scientist at SLB Poland, within the Data Science & AI Hubs. She holds a Master of Science in Mathematics from the Jagiellonian University of Kraków, Poland. Her main research interests are in the fields of machine learning, artificial intelligence, data mining and PHM.

Ahmed Mosallam is the Data Science & AI European Hub Manager at SLB technology center in Clamart, France. He has his Ph.D. degree in automatic control in the field of PHM from University of Franche-Comté in Besançon, France. His main research interests are signal processing, data mining, machine learning and PHM.

Nannan Shen is a Service Quality engineer at SLB with expertise in reliability and asset performance management. She earned her Bachelor's and Master's degrees in biochemistry from Shanghai Jiaotong University. Beginning her career as a Field Engineer in Aberdeen, UK, she transitioned to roles including Operation Support Engineer before assuming her current position in Clamart, SRPC, SLB. Her interests include equipment efficiency, condition based maintenance, PHM and machine leaning.

Fares Ben Youssef is the Reliability and COSD Manager, where he leads a team of Engineers and Technicians focused on several aspects of Well Construction for Offshore Atlantic basin. His responsibilities include delivering Fit for Basin projects, which involve digital, material, mechanical, and electrical design changes tailored to the Offshore Atlantic basin requirements. In 2011, Fares earned a Master's degree in Electronics and Telecommunication Engineering from the University of Paris-Saclay, France.

Defect Data Augmentation Method for Robust Image-based Product Inspection

Youngwoon Choi[1], Hyunseok Lee1, and Sang Won Lee[2]

[1] *Department of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
woonathome@g.skku.edu
ddsa2210@g.skku.edu

[2] *School of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
sangwonl@skku.edu

ABSTRACT

In this paper, we develop a model for detecting defects in fabric products based on an object segmentation algorithm, including a novel image data augmentation method to enhance the robustness. First, a vision-based inspection system is established to collect image data of the fabric products. The three types of fabric defects, such as a hole, a stain, and a dyeing defect, are considered. To enhance defect detection accuracy and robustness, a novel image data augmentation method, referred to as the defect-area cut-mix, is proposed. In this method, the shapes that are the same as each defect are extracted using the masks, and then they are added to non-defective fabric images. Second, an ensemble process is implemented by combining the results of two models, one with high sensitivity in defect diagnosis and the other with lower sensitivity. The results demonstrated that the model trained on the augmented dataset exhibits improved metrics such as intersection over union and classification accuracy in defect detection on the test dataset.

1. INTRODUCTION

Recently, there has been a surge in demand for automation in the product manufacturing and inspection processes across various sectors within the manufacturing industry. However, due to the considerable time and cost, product inspection by small and medium manufacturers is usually done manually with the human eye. Meanwhile, human inspections can lead to inconsistent test results based on the examiner's skill level and fatigue. Recently, thanks to artificial intelligence (AI) technology, manufacturing companies have actively applied automated inspection processes that can be adapted to different types of products (Jung et al., 2021). In order to conduct automated inspections of products with diverse geometries, it is necessary to develop a robust image-based inspection algorithm and an imaging system that is robust to external factors, including lighting conditions.

Research on image-based product inspection algorithms can be divided into many approaches: statistical approach, AI model-based approach and hybrid approach (Hanbay et al., 2016). Among various approaches, the statistical approach involves using image processing techniques (such as frequency decomposition or filtering) to extract features from images, while the hybrid approach combines statistical methods and modeling techniques to leverage the strengths of both. Therefore, these can be recategorized as follows: those based on combined image processing and those based on deep learning models such as convolutional neural networks (CNN) (Bhatt et al., 2021). For combined image processing algorithms that involve a mix of image processing steps, it is possible to achieve high inspection accuracy only for specific products with the same geometries. However, this approach has a drawback: if the product type or capturing environment changes, we must adjust the algorithm's parameters or create a new algorithm from scratch. On the other hand, algorithms based on deep learning models can respond to various products and environments depending on the training dataset and can achieve high inspection accuracy. However, since the performance of the inspection model greatly depends on the quality of the dataset (number of data points, diversity, etc.), developing a robust model requires investing time and effort to collect a large number of images of defective products (Russakovsky et al., 2015).

Combined image processing algorithms are primarily used when the shooting environment is consistent and the variety of inspected products is limited. Tong et al. (2016) presented an optimal Gabor filter for inspecting woven fabrics. Zhou (2019) focused on inspecting defects in semiconductor wafers. They applied a median denoising process to images to extract favorable features for defect detection and proposed algorithms using machine learning techniques such as KNN and SVM to classify images. Deep learning model-based algorithms require significant computing resources and longer processing times compared to combined image processing algorithms. However, they offer the advantage of creating robust models that are resilient to variations in

shooting environments and product types, depending on the quality of the training dataset. Tabernik et al. (2020) presented a novel segmentation network and decision network trained on the Kolektor Surface-Defect-Dataset (KolektorSDD), containing product surface defects, to infer surface defects with higher accuracy. Ho et al. (2021) proposed a step-by-step algorithm where an object detection model locates defects, and an instance segmentation model infers the shape of defects in woven fabrics. Deep learning model-based algorithms are generally evaluated to have higher robustness compared to combined image processing algorithms.

However, if the feature distribution of the images in the training dataset (such as brightness, types of products, defects, etc.) is not sufficiently diverse, the model may only accurately infer data within the feature distribution of the training data. For instance, assuming a specific defect in the training dataset has a radius of 0.1~0.2mm and appears darker than its surroundings, the detection may become challenging if the defect is larger than 0.2mm in radius or if the product's color is darker than the training images. This is due to the new data displaying features not seen in the training. To address this issue, collecting more data would be one solution. However, as mentioned earlier, for small and medium-sized enterprises, investing significant time and resources without immediate productivity gains can be challenging. Therefore, there is a need to generate unseen data using observed feature distributions within the dataset.

Therefore, in this work, we present a novel defect data augmentation method, referred to as the defect-area cut-mix, to improve the accuracy and robustness of deep learning-based fabric defect detection models from the perspective of dataset quality. In addition, an ensemble process is applied by combining high and low sensitivity models in the fabric defect diagnosis. Figure 1 shows the schematic diagram depicting the research methodology in this paper.

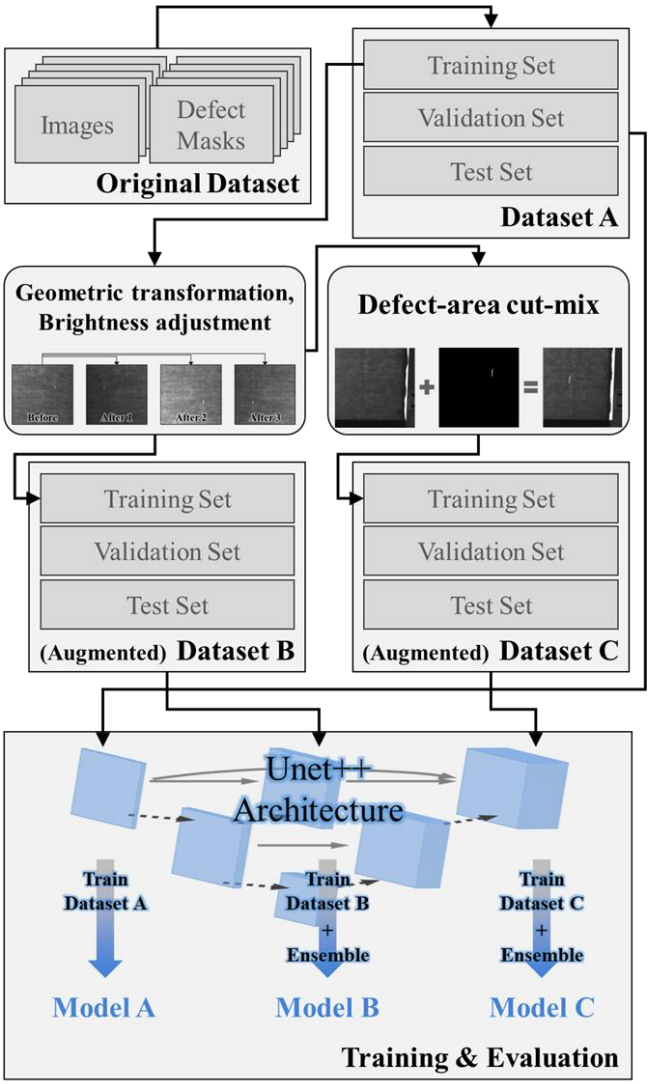


Figure 1. Schematic overview of the research methodology

2. MODEL CONSTRUCTION AND DATA AUGMENTATION

2.1 Image Data Collection and Model Construction

In this section, the research methodology for image collection and defect detection model construction is explained. There are three steps, as follows:

Step 1. Collection of fabric defect image data using vision cameras in the fabric inspection machine.

To collect the fabric defect image data, two single-channel machine vision cameras were installed on the fabric inspection machine, as shown in Figure 2. The fabric inspection machine was designed to inspect the fabric rolls while they were continuously rotating. The three fabric types were black denim, blue denim, and light blue fleece. All images were collected while moving the fabric at a speed of approximately 30 cm per second, like in the actual fabric inspection environment.

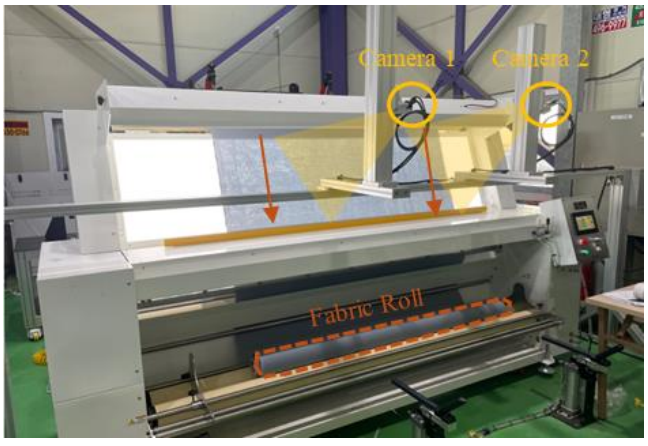


Figure 2. Image data collection using fabric inspection machine

Step 2. Construction of the original datasets and defect-augmented datasets through image patching and masking of defect areas. (Masking and labeling are conducted using CVAT, which is a well-known open-source annotation tool.) (OpenCV et al. n.d.)

The collected original images have dimensions of 3000 pixels in height and 4080 pixels in width. Due to their large size, it is inefficient for deep learning models to process them directly, necessitating resizing or patching. When resizing the original images directly, as shown in Figure 3, small defect areas may become significantly reduced and thus may not be detected. To avoid this issue, we performed patching in a grid of 3 (vertical) by 4 (horizontal) patches and resized each patch to 320x320 pixels. Subsequently, masking was applied to all patches containing defects to construct the dataset.

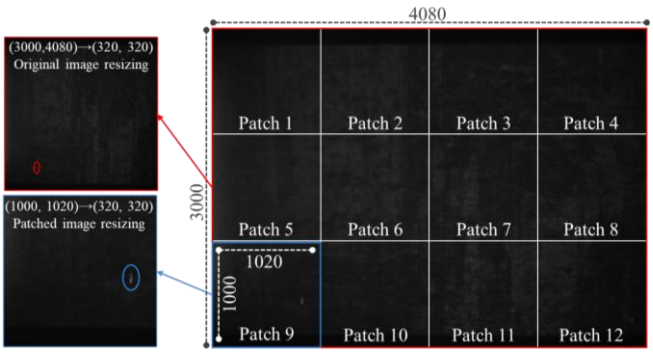


Figure 3. Image patching and resizing

Step 3. Design and construction of the defect detection model architecture for training.

This dataset was then divided into training, validation, and testing sets, as shown in Table 1. Both the training and validation datasets were composed entirely of patches containing defects. For the testing dataset, we included both patches with defects and patches without defects to assess the tendency for false positives in detecting defects on normal (non-defective) patches.

Table 1. The number of images of original patched dataset

Defect type	Without defect	Hole	Stain	Dyeing	Total
Training	0	468	109	71	648
Validation	0	465	95	95	655
Testing	6312	80	45	3	6440

Deep learning-based defect inspection commonly uses object detection and instance or semantic segmentation models. In this study, we selected a semantic segmentation model to reflect the characteristic of quantitatively calculating the area of defects during the quality assessment of fabric products.

We designed our own Unet++ architecture, which has recently demonstrated strong segmentation performance in the biomedical field, to build the model as depicted in Figure 4. The model specifically takes an image input size (320,320) and outputs masks (320,320,4) corresponding to non-defective regions and each defect. Additionally, the input image undergoes four rounds of down-sampling and up-sampling, with skip connections applied between all feature maps (Zhou et al. 2019). To ensure performance in both pixelwise classification and defect classification during model training, we employed the BCE Dice loss function, calculated as Eq. (1).

By using both BCE (Binary Cross-Entropy) and DICE loss functions, it is possible to accommodate diversity in loss calculations while leveraging the stability provided by BCE. In semantic segmentation tasks, x_n and y_n are both binary images (masks), representing the ground truth and the predicted mask, respectively.

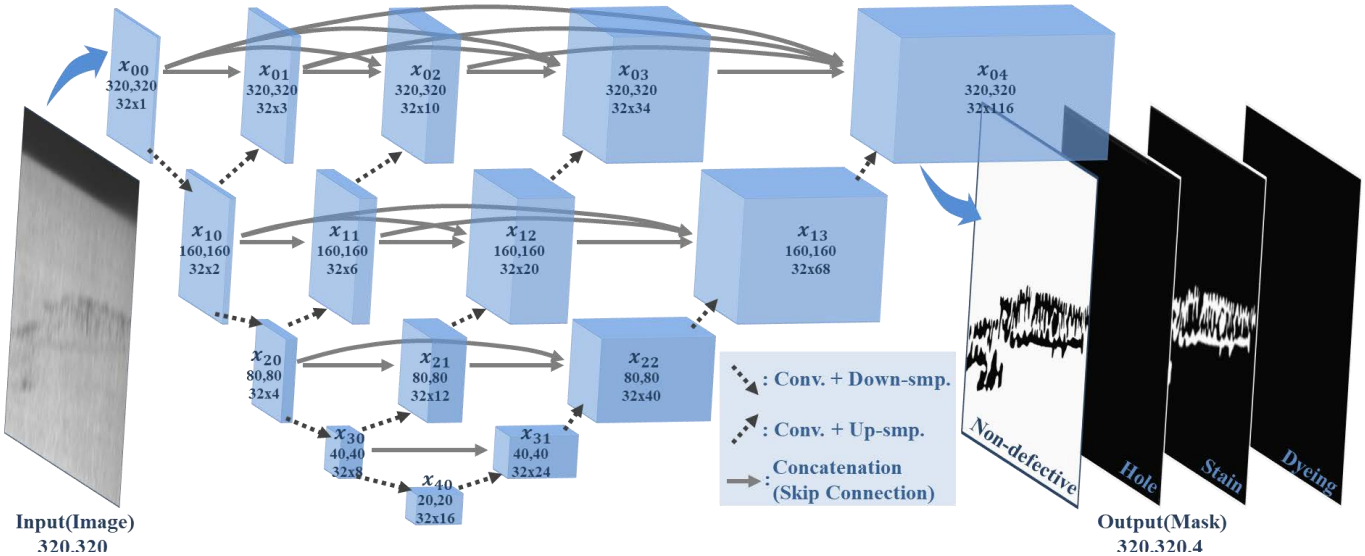


Figure 4. Architecture of fabric defect detection model.

$$\text{BCE Dice Loss} = \text{BCE loss} + 0.5 \times \text{Dice loss}$$

$$\text{BCE Loss} = -\sum(y_n \log \sigma(x_n) + (1 - y_n) \log(1 - \sigma(x_n))) \quad (1)$$

$$\text{Dice loss} = \sum\left(1 - \frac{2 \times y_n \times x_n}{x_n^2 + y_n^2 + 10^{-5}}\right)$$

2.2 Defect Data Augmentation

The number of images within a dataset to ensure the basic performance of CNN-based deep learning models can vary depending on several factors (Luo et al., 2018; Israel et al., 2021). However, models used in industrial applications often face limitations on the available data within their own manufacturing environments. Thus, transfer learning using pre-trained weights on custom datasets is a commonly used strategy to ensure robustness (Redmon et al. 2016). However, even in such cases, it is generally recommended to have at least 1000 images per class (Cho et al., 2015). Therefore, it is necessary to apply a proper data augmentation approach to ensure enhanced performance of the defect detection model.

In the dataset created by patching the collected original images in this study, the numbers of defect instances are highly imbalanced, as shown in Table 1. Additionally, the absolute number of images is insufficient. The most frequently observed defect in both the training and validation datasets is "hole," occurring more than four times as frequently as "stain" and "dyeing" defects combined. Even when combining the training and validation datasets, the total number of images does not exceed 1500. Therefore, image data augmentation is essential, and it is necessary to reflect the characteristics of each defect in the augmentation process. Stain defects exhibit various shapes and intensities within the defect areas, while dyeing defects primarily appear as broad horizontal stripes. Hole defects typically manifest as long, uniformly shaped vertical openings. Figure 5 shows examples of images depicting each type of defect.

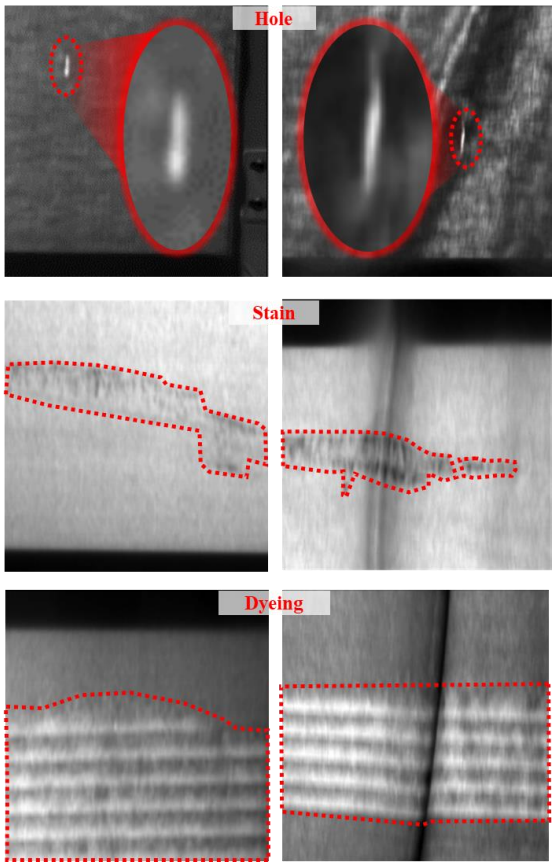


Figure 5. Example images of each type of defects.

Commonly used methods for image augmentation include geometric transformations (such as flipping, rotating, and affine/perspective transformations) and brightness adjustment (such as histogram equalization). These methods are advantageous because the operations applied to the

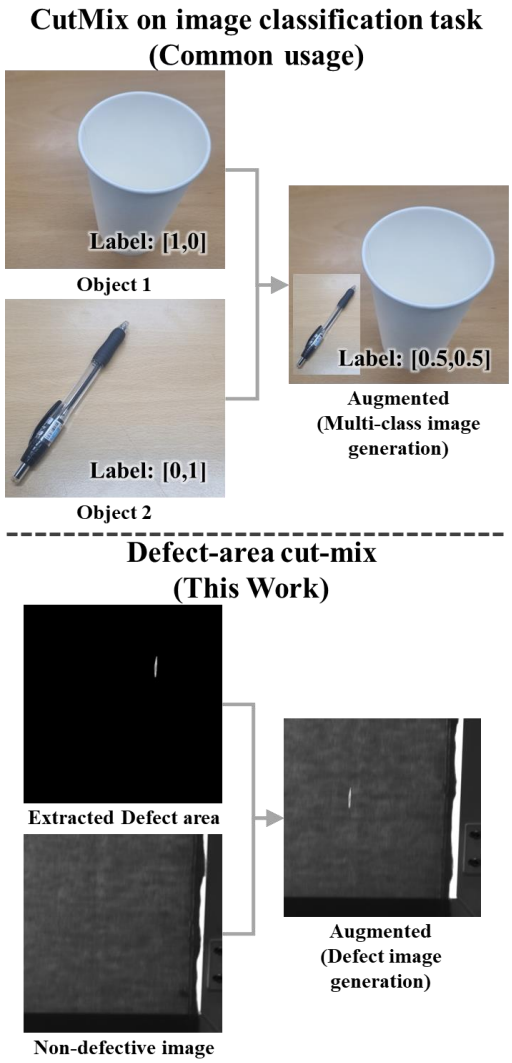


Figure 6. Comparison of conventional cut-mix and defect-area cut-mix.

images are simple, allowing for the rapid generation of new images. However, these methods alter the shape and brightness information of defects along with the surrounding background area, making it challenging to reflect cases where the model inputs unseen shapes of defects.

Therefore, in this study, to overcome the limitations of such data transformations, the cut-mix augmentation technique was applied specifically to defect areas. The cut-mix augmentation is a technique proposed to improve the generalization performance of image classification models by encouraging the model to better learn the local features of each class (Yun et al., 2019). Figure 6 illustrates the difference between the conventional cut-mix augmentation method and the approach used in this study, referred to as "defect-area cut-mix." While the original cut-mix augmentation mixes rectangular regions of different class images directly, in this study, only the defect areas were cut out and overlaid onto images without defects to create new images with defects. This approach allows for the assumption of situations where the shape and location of defects may change from the dataset's perspective.

– Extract defect areas using defect masks and images.

– Select images without defects and overlay defect areas directly or randomly, depending on the characteristics of the defects. Here, the characteristics of the defects refer to the area or shape of the defect areas.

Specifically, for hole defects (see Figure 7), where the defect size is relatively small, the defect areas can be randomly overlaid in portions of the fabric that are not occupied by the fabric. However, manual overlay position selection is required for stain and dyeing defects, where the defect regions' forms and areas vary considerably and may occupy a significant portion of the image.

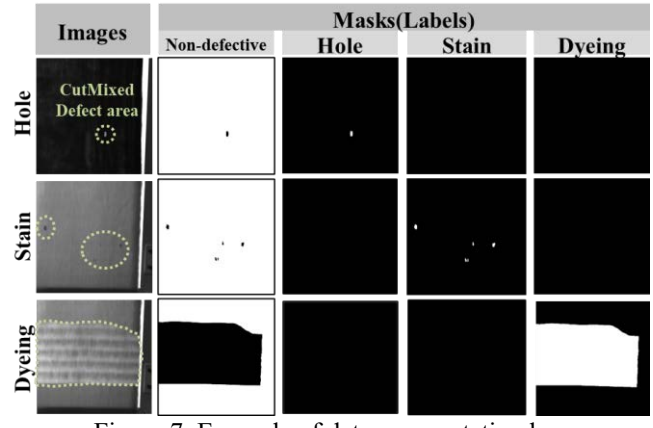


Figure 7. Example of data augmentation by defect-area cut-mix.

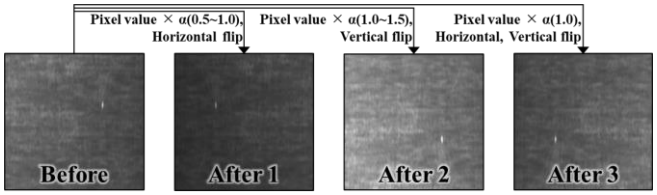


Figure 8. Example of data augmentation by geometric transformation and brightness adjustment.

After applying the defect-area cut-mix as described above, geometric and brightness adjustments were applied to the dataset. The following three image processing techniques were independently applied with a certain probability: brightness adjustment to reflect changes in external lighting conditions during image capture (random alpha values within the range of 0.5 to 1.5 multiplied by the entire image), horizontal flip, and vertical flip. Figure 8 illustrates examples of data augmentation.

To assess the performance difference caused by augmentation, the composition of the final datasets constructed by applying augmentations to the original patched dataset in Table 1 is presented in Table 2 and Table 3 (datasets B and C). To validate the effectiveness of defect-area cut-mix augmentation, augmentations depicted in Figure 7 without defect-area cut-mix were applied to dataset B. On the other hand, every augmentation containing geometric transformation, brightness adjustment, and defect-area cut-mix was applied to dataset C. All augmentations were applied only to the training dataset. However, for stain defects, it was empirically confirmed that applying the defect-area cut-mix technique was not effective due to the diverse shapes of the defects. Therefore, this augmentation was applied to the hole and dyeing defect only. Additionally, to avoid misdiagnosing defects on images without defects, images without defects and with temporary wrinkles (which are not considered defects) were added to the training and validation datasets. Compared to the original patched dataset, the augmented datasets partially alleviated the class imbalance between defect types.

Table 2. Composition of the augmented dataset B

Defect type	Without defect	Hole	Stain	Dyeing	Total
Training	930	549	491	186	2156
Validation	7200	465	95	95	7855
Testing	6312	80	45	3	6440

Table 3. Composition of the augmented dataset C

Defect type	Without defect	Hole	Stain	Dyeing	Total
Training	930	702	907	314	2853
Validation	7200	465	95	95	7855
Testing	6312	80	45	3	6440

3. PERFORMANCE EVALUATION

Through a series of steps, three datasets were constructed before and after augmentation (naturally, the original patched dataset is a subset of the augmented dataset). Each of the three different datasets was trained on our self-designed architecture of the Unet++ model, as depicted in Figure 3. For convenience in description, the model trained on the pre-augmented data will be referred to as Model A, the model trained on the augmented dataset without defect-area cut-mix will be referred to as Model B, and the one trained on the augmented dataset containing defect-area cut-mix will be referred to as Model C. By training a strict model and a less strict model independently, and then using an ensemble method, we were able to avoid incorrectly classifying defect-free fabric data as defective in Model B and Model C. To produce the segmentation results, the ensemble applied to Model B and Model C used soft voting on the pixel-wise classification probabilities of two distinct models.

The identical testing dataset was used to test Models A, B, and C. Masks for the defect and non-defective areas were inferred by the models. The mean Intersection over Union (IOU) score for each patch of each image was used to assess the performance metrics for defect identification, with a 0.7 threshold. This measure evaluates each model's ability to identify defects and distinguish defective regions from non-defective ones.

Furthermore, the confusion matrix depicted in Figure 8 was used to calculate the models' precision, recall, accuracy, and F1 score in order to assess their performance in defect identification. Note that false positives represent instances where non-defective regions were incorrectly predicted as defective, while false negatives represent instances where defects were present but not detected. These metrics provide insights into the classification performance of the models in both non-defective and defective regions. Classification among different types of defects showed a 100% accuracy for all models. This high accuracy can be attributed to the clear characteristics of each type of defect, as described earlier.

Overall performance metrics for Model A, B and C are presented in Table 4. Model C, trained on the augmented dataset with defect-area cut-mix, demonstrated superior performance across all metrics compared to Model A and Model B. Among the performance metrics for defect detection, Recall exhibited the smallest change, indicating that both models excelled at detecting actual defects with minimal variation.

The effects of geometric transformation and brightness adjustment augmentation, as well as additional training on defect-free images, were evident in the difference between models A and B. Comparing the confusion matrix of Model B with that of Model A shows that all metrics improved. Notably, the total number of false positives decreased significantly from 1252 to 181. This indicates a substantial

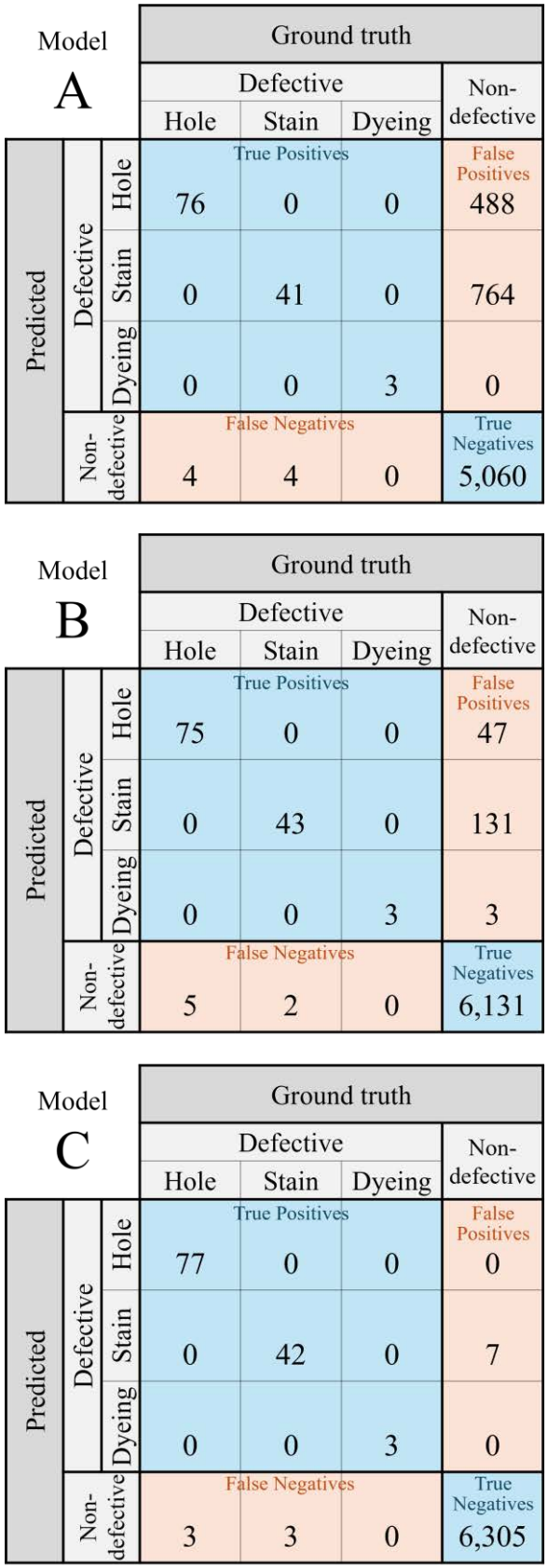


Figure 8. Confusion matrices of Model A, B and C

reduction in the model's tendency to incorrectly classify normal regions as defects. Moreover, the Mean IOU improved significantly from 0.375 to 0.704.

In the case of Model C, trained on a dataset where defect-area cut-mix was applied, the number of false positives decreased significantly compared to the other two models, with only 7 false positives in the Stain defect category. (It is interesting to note that the cut-mix was not applied to the Stain defect.) Furthermore, Model C achieved a more robust test result, with a mean IoU of 0.902, compared to the other two models.

In summary, the augmentation proposed in this study and ensemble techniques led to a significant improvement in defect detection performance compared to model without these enhancements. This demonstrates that the proposed defect-area cut-mix technique can enhance the robustness of deep learning models in image-based defect detection tasks.

Table 4. Metrics comparison of model A, B and C

Model	A	B	C
Mean IOU	0.375	0.704	0.902
Precision	0.086	0.401	0.946
Recall	0.938	0.945	0.953
F1 score	0.160	0.281	0.950
Accuracy	0.804	0.971	0.998

4. CONCLUSION

This study proposed the novel image data augmentation method, referred to as the defect-area cut-mix, for enhancing defect detection accuracy and robustness of the deep learning-based fabric inspection system. In the defect-area cut-mix method, the defect shapes that are the same as actual fabric defects (hole, stain and dyeing defect) were extracted using the masks, and they were added to the non-defective fabric images for an augmentation. To demonstrate the effectiveness of the proposed defect-area cut-mix augmentation method, three data sets were prepared, such as the original dataset with augmentation (dataset A), that with conventional geometrical augmentation and brightness adjustments (dataset B), and that with defect-area cut-mix, geometrical augmentation, and brightness adjustments (dataset C). Then, the ensemble approach combining the deep-learning models with high and low sensitivity was applied to datasets B and C. Finally, it was found that the fabric defect diagnosis model with the dataset C and ensemble approach showed the best performance in terms of mean IOU, precision, recall, F1 score, and accuracy.

ACKNOWLEDGEMENT

This work was supported by the Technology development Program (S3303060, 2022R1A2C3012900) funded by the Ministry of SMEs and Startups (MSS, Korea) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT).

REFERENCES

Jung, W. K., Kim, D. R., Lee, H., Lee, T. H., Yang, I., Youn, B. D., ... & Ahn, S. H. (2021). Appropriate smart factory for SMEs: concept, application and perspective. *International Journal of Precision Engineering and Manufacturing*, *22*, 201-215. DOI 10.1007/s12541-020-00445-2

Hanbay, K., Talu, M. F., & Özgüven, Ö. F. (2016). Fabric defect detection systems and methods—A systematic literature review. *Optik, 127(24)*, 11960-11973. DOI 10.1016/j.ijleo.2016.09.110

Bhatt, P. M., Malhan, R. K., Rajendran, P., Shah, B. C., Thakar, S., Yoon, Y. J., & Gupta, S. K. (2021). Image-based surface defect detection using deep learning: A review. *Journal of Computing and Information Science in Engineering, 21(4)*, 040801. DOI 10.1115/1.4049535

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252. DOI 10.1007/s11263-015-0816-y

Tong, L., Wong, W. K., & Kwong, C. K. (2016). Differential evolution-based optimal Gabor filter model for fabric inspection. *Neurocomputing*, *173*, 1386-1401. DOI 10.1016/j.neucom.2015.09.011

Zhou, Y. (2019). Research on image-based automatic wafer surface defect detection algorithm. *Journal of Image and Graphics*, *7*(1), 26-31. DOI 10.18178/joig.7.1.26-31

Tabernik, D., Šela, S., Skvarč, J., & Skočaj, D. (2020). Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, *31*(3), 759-776. DOI 10.1007/s10845-019-01476-x

Ho, C. C., Chou, W. C., & Su, E. (2021). Deep convolutional neural network optimization for defect detection in fabric inspection. *Sensors*, *21*(21), 7074. DOI 10.3390/s21217074

OpenCV. (n.d.). CVAT. Retrieved [2024.05.10], from https://www.cvat.ai/.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, *39*(6), 1856-1867. DOI 10.1109/TMI.2019.2959609

Luo, C., Li, X., Wang, L., He, J., Li, D., & Zhou, J. (2018, November). How does the data set affect cnn-based image classification performance?. In *2018 5th international conference on systems and informatics (ICSAI)* (pp. 361-366). IEEE. DOI 10.1109/ICSAI.2018.8599448

Israel, I. M., Israel, S. A., & Irvine, J. M. (2021, October). Factors influencing CNN performance. In 2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1-4). IEEE.
DOI 10.1109/AIPR52630.2021.9762112

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*. https://arxiv.org/abs/1511.06348
DOI 10.48550/arXiv.1511.06348

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6023-6032).

BIOGRAPHIES

Youngwoon Choi is now Ph.D. candidate in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree (Mechanical Engineering) from Sungkyunkwan University in 2020 and completed his master's degree in 2022. His research interest is vision AI-based prognostics and health management for smart manufacturing.

Hyunseok Lee is now master candidate in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree (Mechanical Engineering) from Sungkyunkwan University in 2023. His research interest is vision AI-based prognostics and health management for smart manufacturing.

Sang Won Lee is now professor in the school of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree in 1995 and master's degree in 1997 (Mechanical Design and Production Engineering) from Seoul National University. He obtained his Ph.D. degree (Mechanical Engineering) from University of Michigan in 2004. His research interest includes prognostics and health management (PHM), cyber-physical system (CPS), additive manufacturing, and data-driven design.

Detection of Abnormal Conditions in Electro-Mechanical Actuators by Physics-Informed Long Short-term Memory Networks

Chenyang Lai[1], Piero Baraldi[2], Gaetano Quattrocchi[3], Matteo Davide Lorenzo Dalla Vedova[4], Leonardo Baldo[5], Matteo Bertone[6], Enrico Zio[7]

[1,2,7]*Energy Department, Politecnico di Milano, Milano, 20156, Italy*

chenyang.lai@polimi.it
piero.baraldi@polimi.it
enrico.zio@polimi.it

[3,4,5,6]*Department of Mechanics and Aerospace, Politecnico di Torino, Torino, 10129, Italy*

gaetano.quattrocchi@polito.it
matteo.dallavedova@polito.it
leonardo.baldo@polito.it
matteo.bertone@studenti.polito.it

[7] *MINES Paris-PSL University, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, 06964, France*

enrico.zio@mines-paristech.fr

ABSTRACT

Electro-Mechanical Actuators (EMAs) are projected to revolutionize the flight control actuator paradigm, potentially replacing hydraulic-powered systems in the future. Consequently, the functioning of EMAs is destined to become critical for the safe and reliable operation of aircraft. Abnormal conditions of the mechanical components of EMAs can lead to their failure. The objective of this work is to develop a method for the early detection of abnormal conditions of the components of EMAs. The proposed method is based on a signal reconstruction model that estimates the motor position of EMA as expected in normal conditions of its components. Then, the presence of an abnormal condition is identified when the difference between the motor position and its reconstructed position in normal conditions exceeds a preset threshold. The signal reconstruction model employs a Physics-Informed Long Short-Term Memory network (PILSTM), whose architecture combines a physics-informed cell for the solution of the differential equations governing the EMA operation, and a data-driven Long Short-Term Memory (LSTM) cell which receives in input the output of the physics-informed cell and reconstructs the position expected in normal conditions. The proposed method is applied to data simulated by a high-fidelity model of EMAs. The results show that PILSTM is able to provide accurate, physics-consistent estimates of the motor position of EMA in normal conditions and the missed and false detection alarms are lower than those of other state-of-the-art methods.

1. INTRODUCTION

In Prognostics and Health Management (PHM), fault detection amounts to the identification of abnormal conditions in the monitored structure, system and components (SSCs). A common approach relies on signal reconstruction models that give in output the signal values in normal conditions of the SSC (Hines, Uhrig, & Wrest, 1998). The difference between the actual signal measurements and the reconstructed signal values (so-called residual) is analyzed for detecting the presence of abnormal conditions: the larger the residuals, the more the SSC behaviour deviates from that in normal conditions.

Signal reconstruction methods can be classified in model-based and data-driven (Yang, Ling, & Bingham, 2013). Model-based approaches typically use numerical simulators which code the specific laws of physics. They require a limited amount of data for model parameter calibration and retain the physical interpretability of the model output. In (Zhang, Foo, Don Vilathgamuwa, Tseng, Bhangu, & Gajanayake, 2013), a method combining physics-based model and extended Kalman filter has been developed to perform fault detection of induction motors. In (Sarikhani, & Mohammed, 2012), a back electromotive force estimator has been built using only laws of physics and the fault detection is performed by comparing the signal estimates and the

nominal values. The main limitations of model-based approaches are the difficulty of representing with a model the complexity of modern SSCs, the inevitable approximations made to build and solve the model and the computational efforts needed for numerical simulation. On the contrary, data-driven methods are typically simpler to implement since they do not require knowledge on the physics of the system and are able to discover complex nonlinear patterns in data. Recently, deep learning-based methods have gained popularity. In (Qi, Jang, Cui, & Moon, 2023), a data-driven model based on the use of gated recurrent units has been built to reconstruct the dynamic behaviour of stirred tank reactors in normal conditions. In (Xu, Baraldi, Lu, & Zio, 2022), a Generative Adversarial Network and an auxiliary encoder have been developed to detect anomalies in the operation of automatic doors of high-speed trains. Some limitations of data-driven methods are their difficulty of extrapolation outside the region covered by the training data and the lack of interpretability of their outputs, due to their black-box nature.

Hybrid methods combining model-based and data-driven methods have also been proposed. In (Chao, Kulkarni, Goebel, & Fink, 2022), sensor readings and estimates of unobservable parameters inferred by physics-based models have been used as input of a deep learning method for predicting the remaining useful life of turbofan engines. In (Shen, Lu, Sadoughi, Hu, Nemani, Thelen, ... & Kenny, 2021), a physics-informed deep learning approach has been developed for fault detection of bearings, in which the loss function contains a term that incorporates physical knowledge on the envelop spectrum. In (Yucesan, & Viana, 2021), recurrent neural networks embedded with physics-based models of fatigue and grease degradation have been developed to predict grease damage. In (Li, & Deka, 2021), a physics-informed autoencoder integrating the laws of physics relating current and voltage in the loss function is developed to detect high impedance faults in distribution grids. In (Chen, Rao, Feng, & Zuo, 2022), a physics-informed strategy for setting the hyperparameters of a Long Short-Term Memory (LSTM) network is developed for fault detection of gearboxes. It is based on the maximization of the discrepancy between healthy and simulated faulty patterns. Physics-informed methods have shown their capability of improving performance in fault detection, diagnostics and prognostics, while enhancing consistency with the law of physics, which can enhance trustability. Yet, their use for developing signal reconstruction models is still challenged by the difficulty of considering variable operating conditions and highly nonlinear relationships.

In this context, this work presents the development of a novel signal reconstruction method based on the use of Physics-informed LSTMs (PILSTMs) to perform fault detection. The basic idea behind the developed PILSTM is the combination of a physics-informed and a data-driven layers. The former solves the differential equations of the model of the system in normal conditions, whereas the latter layer receives in input the output of the physics-informed layer and reconstructs the signals in normal conditions. The developed approach is applied to data simulated using a high-fidelity model of EMAs, and its performance is compared to those of other state-of-the-art methods. The problem of fault detection in EMAs has been previously addressed in (Yang, Guo, & Zhao, 2019), where a LSTM-based model is developed for signal prediction and the residuals between predictions and measurements are used to detect abnormal conditions. In (Zhang, Tang, & Chen, 2021), a model based on an improved Gate Recurrent Unit (GRU) is developed to predict signal evolutions, which are then used to classify faults with a similarity measure. The two methods have been developed and verified considering a small set working conditions and command signals, which limits their applications to the real scenarios.

The rest of this paper is organized as follows. Section 2 formulates the problem. Section 3 presents the proposed fault detection method. Section 4 discusses the application of the proposed method to EMAs. Finally, the conclusions of the work are presented in Section 5.

2. PROBLEM STATEMENT

We consider the motor of an EMA, which is its most critical component, considering the frequency of its failures and their potential severity. The function of the motor is to provide the torque needed to actuate the aircraft aerodynamic surface (Berri, Dalla Vedova, & Maggiore, 2019). It is here modelled as a component that receives in input the three-phase current signals, $[x_{\tau,1}, x_{\tau,2}, x_{\tau,3}]$, and provides as output the motor position, y_τ (Baldo, Bertone, Dalla Vedova, & Maggiore, 2022) (Figure 1).

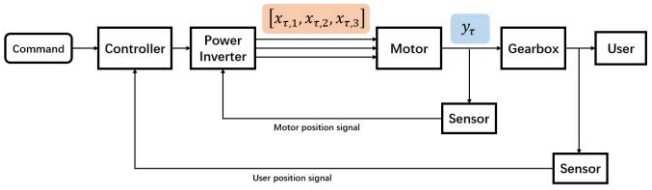


Figure 1. Input and output signals of the motor of an EMA.

We assume to have the available dataset $\mathcal{D}^{train} = \{X^r, y^r\}_{r=1,\dots,R}$ containing R input-output time-series of T time instants, $X^r \in \mathbb{R}^{T\times 3}$ and $y^r \in \mathbb{R}^T$, collected during the operation of an EMA in normal conditions. The generic vector $\vec{x}_t^r = [x_{t,1}, x_{t,2}, x_{t,3}]$ of X^r contains the value of the nth phase current signal at the tth time instant, whereas y_t^r indicates the value of the motor position signal. Each input-output time series $\{X^r, y^r\}$ corresponds to different operational conditions of the EMA.

The objective is the development of a fault detection method for the early identification of abnormal conditions in the EMA motor. The method is based on the development of a

signal reconstruction model $f: \hat{y}_\tau = f(X_\tau)$ that gives in output the value $\hat{y}_\tau$ that the motor position would have in normal conditions at the time τ, given the values of the input signals $X_\tau = [\vec{x}_t]_{t=1:\tau}$ measured from time 0 until the time τ. An anomaly indicator is, then, built considering the residual, $d_\tau = \hat{y}_\tau - y_\tau$, between the motor position reconstructed by the model and the actual measurement. The norm of d_τ is small if EMA is in normal conditions and large in case of abnormal conditions: therefore, the detection of an anomaly condition can be obtained by statistical analysis of the residuals.

3. METHOD

The signal reconstruction model, $\hat{y}_\tau = f(X_\tau)$ is a PILSTM, which is capable of dealing with large non-linearities in the dynamics of the time-series. The first layer is a physics-informed (PI) layer that implements numerical methods to estimate the motor position. The second layer is a traditional LSTM cell. Fully-connected (FC) layers are used to map the extracted hidden features to the output signal (motor position). The complete architecture of the signal reconstruction model is shown in Figure 2.

Section 3.1 describes the PI layer and Section 3.2 describes the LSTM layer of the PILSTM. Section 3.3 defines the anomaly indicator used for the detection of abnormal conditions.

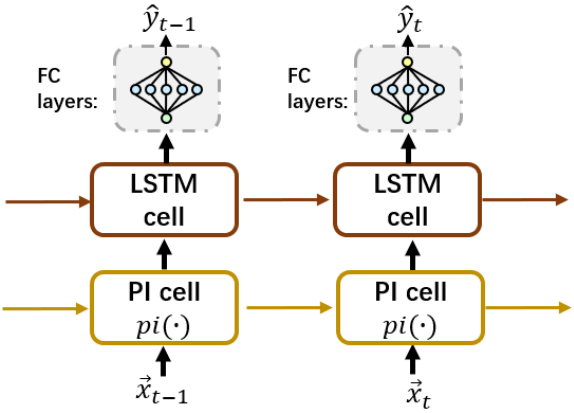


Figure 2. Architecture of the proposed PILSTM.

3.1. Physics-informed layer

The differential equation used to describe the operation of the EMA motor and transmission in normal conditions is (Baldo, Bertone, Dalla Vedova, & Maggiore, 2022):

$$F = (J_m + J_u^*)\ddot{y}_t + C_u\dot{y}_t + C_m(\pm\sqrt{|\dot{y}_t|}) \quad (1)$$

where y_t is the motor position, F is the motor torque, J_m is the motor inertia, J_u^* is the inertia of the gearbox following the motor and C_u is the viscous friction of the gearbox. F is computed from the current signals $x_{t,1}, x_{t,2}$ and $x_{t,3}$ as:

$$F = \sum_{n=1,2,3} x_{t,n} \cdot k_n \quad (2)$$

with

$$k_1 = -k_E \cdot \sin(\theta_e) \quad (3)$$

$$k_2 = -k_E \cdot sin\left(\theta_e - \frac{2}{3}\pi\right) \quad (4)$$

$$k_3 = -k_E \cdot sin\left(\theta_e - \frac{4}{3}\pi\right) \quad (5)$$

$$\theta_e = 2\pi\left(\frac{P \cdot y_t}{2\pi} - floor\left(\frac{P \cdot y_t}{2\pi}\right)\right) \quad (6)$$

where k_n is the nth-phase back-electromotive force (EMF) coefficient, P is the number of pole pairs and k_E is the back-EMF motor constant.

The EMA motion in normal conditions is, then, formulated as a 2-order differential equation:

$$\ddot{y}_t = g(\vec{x}_t, \dot{y}_t, y_t) \quad (7)$$

which is numerically solved by resorting to the 4-stage Runge–Kutta method (RK4) (Butcher, 1987). More details about the RK4 method are reported in Appendix 1. The customized RK4 cell solves Eq (1) by computing (Nascimento, Fricke, & Viana, 2020):

$$[y_t, \dot{y}_t] = pi(\vec{x}_t, y_{t-1}, \dot{y}_{t-1}) \quad (8)$$

The obtained y and $\dot{y}$ are hidden features $\vec{h}^{(1)}$ fed to the LSTM layer. To distinguish the estimates provided by the physics-informed layer and the actual measurements, the motor position and its first derivative computed by the physics-informed layer are indicated as y^{phy} and $\dot{y}^{phy}$ (Figure 3).

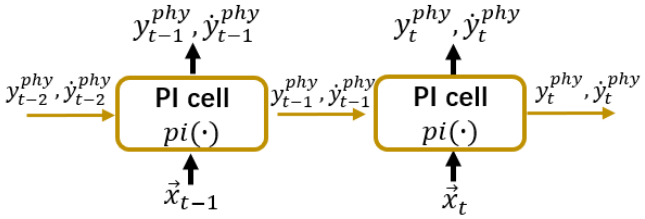


Figure 3. Architecture of the physics-informed layer.

3.2. LSTM layer

In the second layer of the PILSTM, data-driven LSTM cells are used to reconstruct the signal values. The cells receive in input the signal estimates $\vec{h}^{(1)}$ (Figure 4) and control the information flow using input, forget and output gates to remember, when needed, information for long periods of time.

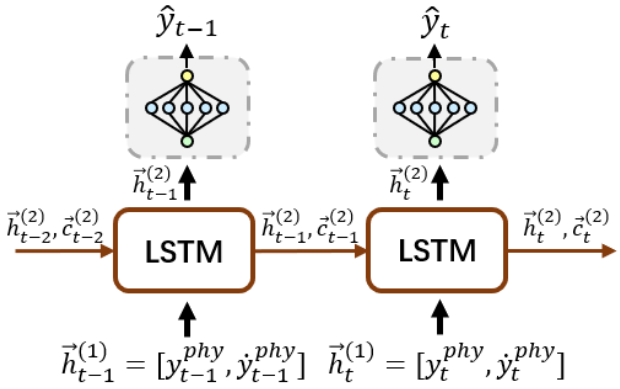


Figure 4. Architecture of LSTM layer.

Specifically, at time t, the LSTM cells receive in input $\mathbf{h}_t^{(1)}$, which is the concatenation of y_t and $\dot{y}_t$, process the temporal behaviour of the time-series and provide in output the vector of hidden features $\vec{h}_t^{(2)}$ and cell states $\vec{c}_t^{(2)}$. The following FC layers map the extracted features $\vec{h}_t$ into the estimates of the motor position $\hat{y}_t$. The structure and detailed operation of the LSTM cells are reported in Appendix 2.

The objective of the training of PILSTM is to identify the optimal combination of parameters values (weights and biases) that minimizes the error between the actual values of y_t and the estimates $\hat{y}_t$. To this aim, the following loss function is minimized on the training data in $\mathcal{D}^{train}$:

$$\mathcal{L} = \frac{1}{R \cdot T} \sum_{r=1}^{R} \sum_{t=1}^{T} \|y_t^r - \hat{y}_t^r\|^2 \qquad (9)$$

3.3. Definition of the anomaly indicator

The test time-series $X^{test} = \{x_{t,n}^{test}\}_{t=1,\dots,\tau,n=1,2,3}$, contains the measurements of the phase current signals until the present time τ, and it is rearranged in a set of $\frac{\tau-l}{ss} + 1$matrices $X_{t_k}^{tw} \in \mathbb{R}^{l\times 3}$, $t_k = (k-1) * ss + l$ with $k = 1,\dots,\frac{\tau-l}{ss} + 1$, each one containing the current signals $[x_{t,1}, x_{t,2}, x_{t,3}]$ in a time window of l time steps. Between one matrix and the following, a sliding step of ss time steps is applied. At present time τ, the residuals $D = [\hat{y}_{\tau-l+1}, \hat{y}_{\tau-l+2}, \dots, \hat{y}_\tau] - [y_{\tau-l+1}, y_{\tau-l+2}, \dots, y_\tau]$ are computed and used to define the anomaly indicator (AIND):

$$AIND = \|D\|_{L_2}^2 \qquad (10)$$

Finally, a threshold Thr for $AIND$ is defined: considering a validation set, and the occurrence of an abnormal condition is detected if $AIND$ exceeds Thr.

4. Case Study

The functioning of an EMA working in normal conditions has been simulated using the high-fidelity (HF) simulator described in (Berri, Dalla Vedova, & Maggiore, 2019). Specifically, $R = 60$ time-series with time length $T = 50$s of EMA operation in normal conditions have been generated for training the PILSTM model and verifying the signal reconstruction performance. Wavelet denoising has been applied to measured signals. The time series has been obtained at a frequency of 100Hz.

4.1. Signal reconstruction in normal conditions

The dimensionality of $\vec{h}_t^{(2)}$ is 10. The FC layers consist of 2 hidden layers with 10 and 5 neurons and 1 output layer with 1 neuron. The learning rate is set equal to 0.01 and the epoch is 250. The Adam optimizer is used to optimize the parameters of the LSTM layer and FC layers.

The $R = 60$ time-series in normal condition are divided into a training set containing 30 time-series and a validation set containing the other 30 time-series.

The performance of the proposed PILSTM is compared to two state-of-the-art methods: (1) a pure physics-based approach based on the solution of Eq. (13) with RK4 to compute the motor position; (2) a pure data-driven method, which uses a traditional LSTM to estimate the motor position. The most critical hyperparameters of the LSTM (number of layers, number of hidden states, learning rate) are optimized performing a grid-search with the objective of maximizing the reconstruction accuracy evaluated on a subset of the training set not used for the loss computation (Eq. (9)) during training. The optimal configuration is found to be 2 layers, 16 hidden states and a learning of 0.001. The Root Mean Squared Error (RMSE) is used as metric to evaluate the reconstruction performance.

Table 1. Comparison of the accuracy in the reconstruction of the motor position for EMA working in normal conditions. The RMSE is computed with respect to the time series of the validation set.

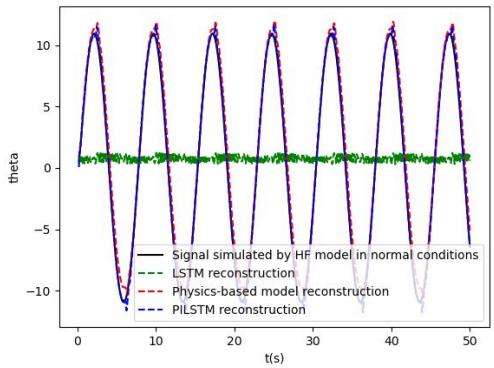
	PILSTM	Physics-based method	LSTM
RMSE	0.7087	0.9621	14.6087

Table 1 reports the obtained reconstruction accuracy on the validation data. An example of motor position estimates is shown in Figure 5. Note that: (1) the pure data-driven method provides the worst performance, which indicates that the training data are not providing enough information for the reconstruction of the input-output relationship; (2) the proposed PILSTM method provides the best performance, which is obtained by reducing the systematic error of the pure physics-based model in the position reconstruction when the motor is operating reversely (Figure 5). Due to the remarkably worst performance of the LSTM model, which makes reconstruction errors more than one order of

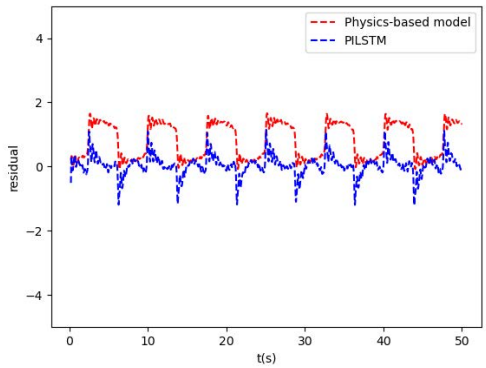
magnitude larger than the other approaches, the generated residuals by LSTM are not shown in Figure 5 (b).

4.2. Fault detection

A test set containing normal and abnormal conditions data has been generated by simulator. The set contains 9 time-series in normal conditions and 20 time-series in abnormal conditions obtained by assuming a dry friction of 20%, 35% and 50%, respectively, for a time interval of 50s. With respect to the anomaly indicator setting, the length of the time window is set equal to 100, which is the number of measurements collected in 1s, and the sliding step ss is set equal to 10 steps. An example of reconstruction of the motor position for an EMA operating in abnormal conditions is shown in Figure 6. As expected, the values of the residuals tend to be larger than the residuals in normal conditions (Figure 5 (b)) for both the physics-based model and the proposed method. Also, the residuals of the proposed method are remarkably larger than zero, which confirms its capability of distinguishing between normal and abnormal conditions.

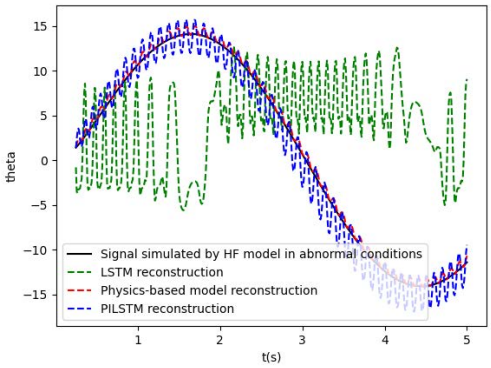


(a)

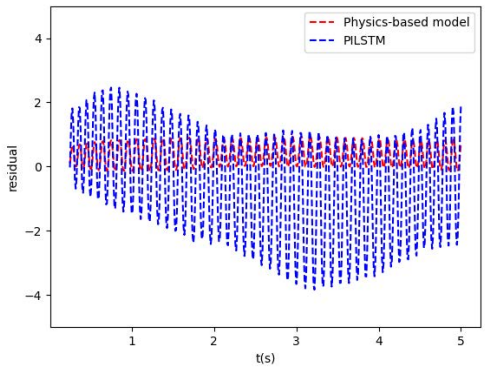


(b)

Figure 5. (a) reconstructions of the motor position and (b) corresponding residuals for an EMA in normal conditions.



(a)



(b)

Figure 6. (a) reconstructions of the motor position and (b) corresponding residuals for an EMA in abnormal conditions.

Figures 7, 8, 9 show the Receiver Operating Characteristic (ROC) curves obtained by varying the threshold of the anomaly indicator for the detection, considering the three levels of fault severity, separately. The x-axis reports the False Positive Rate (FPR), i.e., the rate of time windows in normal conditions identified as abnormal conditions and the y-axis reports the True Positive Rate (TPR), i.e., the rate of time windows in abnormal conditions identified indeed as abnormal conditions. The ideal performance is represented by the upper left corner point [0,1]. An overall measure of anomaly detection performance is the AUC (Area under the ROC Curve) whose most satisfactory value is 1, which indicates that all normal and abnormal time series are correctly identified.

Table 2. AUC considering the three levels of fault severity.

Fault severity	Proposed method: PILSTM	Physics-based method	Pure data-driven method LSTM
20% dry friction	0.7211	0.4590	0.4056

35% dry friction	0.7661	0.4637	0.5468
50% dry friction	0.9971	0.4892	0.4628

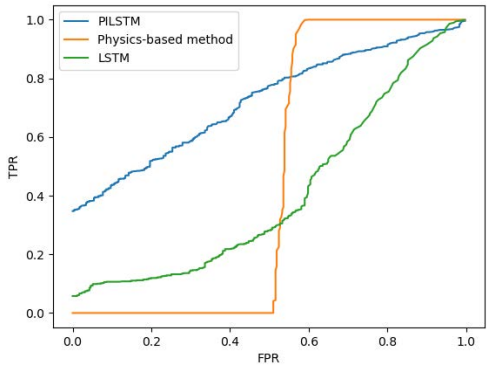


Figure 7. ROC curve made by normal condition data and 20% dry friction abnormal condition data in the test set.

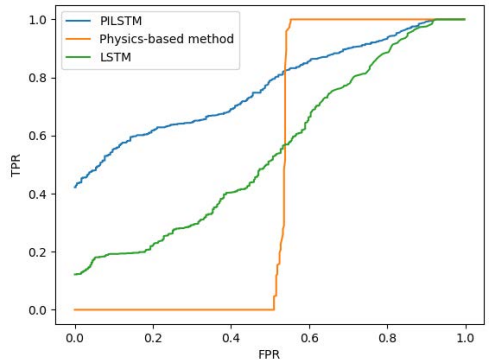


Figure 8. ROC curve made by normal condition data and 35% dry friction abnormal condition data in the test set.

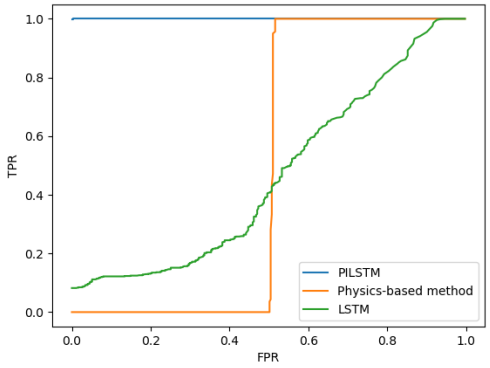


Figure 9. ROC curve made by normal condition data and 50% dry friction abnormal condition data in the test set.

From Table 2, it is seen that the developed PILSTM provides the best performance in all fault severities and the pure LSTM model shows the worst performance due to its poor capability of reconstructing the signals. As expected, the performance of PILSTM becomes better as the fault severity increases.

5. Conclusion

The present work has addressed the problem of fault detection in industrial components. A novel method for signal reconstructions has been developed based on a PILSTM model. Specifically, the PILSTM integrates an RK4 solver of the differential equation governing the EMA operation in normal condition into a LSTM hidden layer. A case study considering a simulated dataset of EMA operation has been considered. The proposed method has shown a more satisfactory accuracy in the signal reconstruction than pure data-driven and physics-based methods. The residuals between reconstructed and measured signals have, then, been used for the detection of abnormal conditions. The results show that the method is capable of detecting abnormal conditions of smaller severity than other comparison methods.

Future work will be devoted to optimally setting the threshold used for detecting the occurrence of abnormal conditions, with the objective of balancing false and missed alarms for fault detection according to the user demand. Also, the obtained results will be compared with those of other state-of-the-art methods for unsupervised abnormal condition detection, such as Deep Semi-supervised Anomaly Detection.

Acknowledgement

Chenyang Lai gratefully acknowledges the financial support from the China Scholarship Council (No. 202006290009). The work of Piero Baraldi is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence). The work of Enrico Zio is supported by iRel40 European co-funded innovation project, granted by the ECSEL Joint Undertaking (JU) under grant agreement No 876659.

References

Baldo, L., Bertone, M., Dalla Vedova, M. D., & Maggiore, P. (2022). High-Fidelity Digital-Twin Validation and Creation of an Experimental Database for Electromechanical Actuators Inclusive of Failures. In *2022 6th International Conference on System Reliability and Safety (ICSRS)* (pp. 19-25). IEEE.

Baraldi, P., Di Maio, F., Genini, D., & Zio, E. (2015). Comparison of data-driven reconstruction methods for fault detection. *IEEE Transactions on Reliability*, *64*(3), 852-860.

Berri, P. C., Dalla Vedova, M. D., & Maggiore, P. (2019). A lumped parameter high fidelity EMA model for model-based prognostics. *Proceedings of the 29th ESREL, Hannover, Germany*, 22-26.

Butcher, J. C. (1987). *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*. Wiley-Interscience.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Chen, Y., Rao, M., Feng, K., & Zuo, M. J. (2022). Physics-Informed LSTM hyperparameters selection for gearbox fault detection. *Mechanical Systems and Signal Processing*, *171*, 108907.

Hines, J. W., Uhrig, R. E., & Wrest, D. J. (1998). Use of autoassociative neural networks for signal validation. *Journal of Intelligent and Robotic Systems*, *21*, 143-154.

Li, W., & Deka, D. (2021). Physics-informed learning for high impedance faults detection. In *2021 IEEE Madrid PowerTech* (pp. 1-6). IEEE.

Nascimento, R. G., Fricke, K., & Viana, F. A. (2020). A tutorial on solving ordinary differential equations using Python and hybrid physics-informed neural network. *Engineering Applications of Artificial Intelligence*, *96*, 103996.

Qi, M., Jang, K., Cui, C., & Moon, I. (2023). Novel control-aware fault detection approach for non-stationary processes via deep learning-based dynamic surrogate modeling. *Process Safety and Environmental Protection*, *172*, 379-394.

Sarikhani, A., & Mohammed, O. A. (2012). Inter-turn fault detection in PM synchronous machines by physics-based back electromotive force estimation. *IEEE Transactions on Industrial Electronics*, *60*(8), 3472-3484.

Shen, S., Lu, H., Sadoughi, M., Hu, C., Nemani, V., Thelen, A., ... & Kenny, S. (2021). A physics-informed deep learning approach for bearing fault detection. *Engineering Applications of Artificial Intelligence*, *103*, 104295.

Xu, M., Baraldi, P., Lu, X., & Zio, E. (2022). Generative Adversarial Networks With AdaBoost Ensemble Learning for Anomaly Detection in High-Speed Train Automatic Doors. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 23408-23421.

Yang, J., Guo, Y., & Zhao, W. (2019). Long short-term memory neural network-based fault detection and isolation for electro-mechanical actuators. *Neurocomputing*, *360*, 85-96.

Yang, Z., Ling, B. W. K., & Bingham, C. (2013). Fault detection and signal reconstruction for increasing operational availability of industrial gas turbines. *Measurement*, *46*(6), 1938-1946.

Yucesan, Y. A., & Viana, F. A. (2021). Hybrid physics-informed neural networks for main bearing fatigue prognosis with visual grease inspection. *Computers in Industry*, *125*, 103386.

Zhang, X., Foo, G., Don Vilathgamuwa, M., Tseng, K. J., Bhangu, B. S., & Gajanayake, C. (2013). Sensor fault detection, isolation and system reconfiguration based on extended Kalman filter for induction motor drives. *IET Electric Power Applications*, *7*(7), 607-617.

Zhang, X., Tang, L., & Chen, J. (2021). Fault diagnosis for electro-mechanical actuators based on STL-HSTA-GRU and SM. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-16.

APPENDIX 1: 4-STAGE RUNGE–KUTTA METHOD

With the available physical knowledge of industrial components, a generalized form of 2-order differential equation that governs the industrial component can be defined:

$$\ddot{y} = g(\vec{x}, \dot{y}, y) \tag{11}$$

where $\vec{x}$ is the vector of input signals, y is the vector output signals, $\dot{y}$ and $\ddot{y}$ are first and second derivatives of y, respectively.

Considering the step-size h between time t and time $t+1$, RK4 is used to numerically integrate Eq. (11) over time with step h:

$$\begin{bmatrix} \dot{y}_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} \dot{y}_t \\ y_t \end{bmatrix} + \frac{h}{6} \cdot \begin{bmatrix} k_1 + 2k_2 + 2k_3 + k_4 \\ l_1 + 2l_2 + 2l_3 + l_4 \end{bmatrix} \tag{12}$$

$$k_1 = g(\vec{x}_t, \dot{y}_t, y_t) \tag{13}$$

$$k_2 = g\left(\vec{x}_{t+h/2}, \dot{y}_t + \frac{h}{2} \cdot k_1, y_t + h \cdot \frac{l_1}{2}\right) \tag{14}$$

$$k_3 = g\left(\vec{x}_{t+h/2}, \dot{y}_t + \frac{h}{2} \cdot k_2, y_t + h \cdot \frac{l_2}{2}\right) \tag{15}$$

$$k_4 = g(\vec{x}_{t+h}, \dot{y}_t + h \cdot k_3, y_t + h \cdot l_3) \tag{16}$$

$$l_1 = y_t \tag{17}$$

$$l_2 = y_t + \frac{h}{2} \cdot l_1 \tag{18}$$

$$l_3 = y_t + \frac{h}{2} \cdot l_2 \tag{19}$$

$$l_4 = y_t + h \cdot l_3 \tag{20}$$

APPENDIX 2: LSTM CELL

LSTM cell structure at time *t* is shown in Figure 9. We differentiate output and states of LSTM cell denoted as $\vec{h}_t$ and $\vec{c}_t$, respectively. Vector size of the output and states is the same and it is defined by number of hidden states in the cell. Let denote p as number of hidden states, so $\vec{h}_t \in \mathbb{R}^{p\times 1}$ and $\vec{c}_t \in \mathbb{R}^{p\times 1}$. The $\vec{h}_{t-1}$ and $\vec{c}_{t-1}$ of LSTM cell at time $t-1$ will serve as an input to LSTM cell at time t, whereas the other input is $\vec{x}_t$. There are three gates that control the information flow within cell: (1) input gate $\vec{i}_t \in \mathbb{R}^{p\times 1}$ controls what information based on output $\vec{h}_{t-1}$ and $\vec{x}_t$ will be passed to memory cell, (2) output gate $\vec{o}_t \in \mathbb{R}^{p\times 1}$ controls what information will be carried to the next time step and (3) forget gate $\vec{f}_t \in \mathbb{R}^{p\times 1}$ controls how memory cell will be

updated. All LSTM cells that are used in the models are implemented as follows:

$$\vec{i}_t = \sigma(W_i\vec{x}_t + U_i\vec{h}_{t-1} + \vec{b}_i) \quad (21)$$

$$\vec{o}_t = \sigma(W_o\vec{x}_t + U_o\vec{h}_{t-1} + \vec{b}_o) \quad (22)$$

$$\vec{f}_t = \sigma(W_f\vec{x}_t + U_f\vec{h}_{t-1} + \vec{b}_f) \quad (23)$$

$$\vec{a}_t = tanh(W_c\vec{x}_t + U_c\vec{h}_{t-1} + \vec{b}_c) \quad (24)$$

$$\vec{c}_t = \vec{f}_t \circ \vec{c}_{t-1} + \vec{i}_t \circ \vec{a}_t \quad (25)$$

$$\vec{h}_t = \vec{o}_t \circ tanh\,(\vec{c}_t) \quad (26)$$

where variable weights and bias to be computed during training process are W_i, W_o, W_f, $W_c \in \mathbb{R}^{p\times L}$, U_i, U_o, U_f, $U_c \in \mathbb{R}^{p\times p}$, $\vec{b}_i$, $\vec{b}_o$, $\vec{b}_f$, $\vec{b}_c \in \mathbb{R}^{p\times 1}$. $\circ$ is element-wise multiplication of two vectors (Hadamard product). σ is element-wise logistic sigmoid activation function. Connection between different layers of LSTMs is achieved such that the output of former layer is as an input to the next layer.

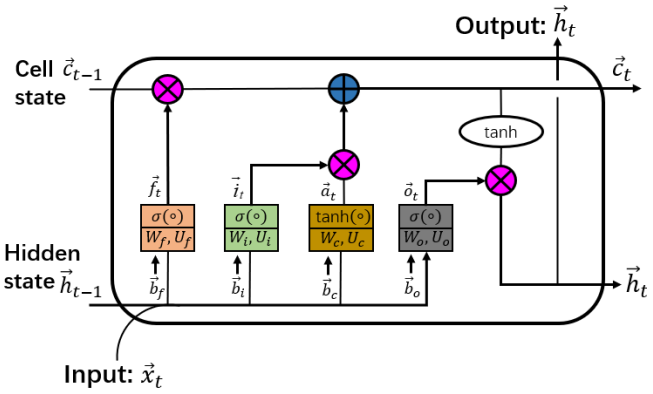


Figure 9. Single LSTM cell structure.

Development of a Feature Extraction Methodology for Prognostic Tasks of Aerospace Structures and Systems

Antonio Orrù[1], Thanos Kontogiannis[2], Francesco Falcetelli[3], Raffaella Di Sante[4], Nick Eleftheroglou[5]

[1,3,4] *Department of Industrial Engineering, University of Bologna, Italy*
antonio.orru@studio.unibo.it
francesco.falcetelli@unibo.it
raffaella.disante@unibo.it

[1,2,5] *Intelligent and Sustainable Prognostics Group, Aerospace Structures and Materials Department, Faculty of Aerospace Engineering, TU Delft, Delft, 2629 HS, Netherlands*
a.kontogiannis@tudelft.nl
n.eleftheroglou@tudelft.nl

ABSTRACT

The performance of prognostic models used for prognostic health management (PHM) applications heavily depend on the quality of features extracted from raw sensor data. Traditionally, feature extraction criteria such as monotonicity, prognosability, and trendability are selected intuitively. However, this intuitive selection may not be optimal.

This research introduces an innovative approach to transform raw data into 'high-scoring' data without the need for predefined feature extraction criteria. Our methodology involves generating a set of synthetic high-scoring time series. These synthetic time series, resembling the length and amplitude of target features, are created through Monte Carlo sampling (MCS) of a user-defined hidden semi-markov model (HSMM). We pair these synthetic time series with raw data/features from the signals and use them as targets to train a convolutional neural network (CNN). As a result, the trained CNN can convert input features into high-scoring ones, irrespective of their initial characteristics. So, this study provides the following contribution to PHM frameworks: it transforms raw data/features into high-scoring ones without relying on predefined criteria, rather on stochastically generated labels that resemble the nature of the degradation processes. It is worth noting, that the proposed FE technique is independent of the prognostic model that will be utilised, thus making it applicable to the established prognostic models.

We demonstrate and validate the effectiveness of this approach using acoustic emission (AE) sensor data for remaining useful life (RUL) estimation of open-hole CFRP specimens. We compare prognostic performance using cumulative AE features with their transformations via our proposed framework. The transformed features exhibit superior prognostic performance, underscoring the value of our innovative feature extraction framework.

1. INTRODUCTION

The current state-of-the-art feature extraction (FE) for prognostics relies heavily on deterministic targets chosen based on intuitively defined metrics such as monotonicity, prognosability, and trendability (Coble & Hines, 2009). These targets have shown efficacy in transforming raw data from sensors, providing a foundational approach for modelling degradation histories (Eleftheroglou, 2020; Moradi, Broer, Chiachío, Benedictus, & Zarouchas, 2023).

The literature-standard procedure for FE for prognostics includes the transformation of the noisy sensor data to high-scoring ones by utilising predefined deterministic labels. These labels are usually derived from simple functions such as second-degree polynomials, exponential and logarithmic. However, the inherent limitation of these deterministic labels lies in their assumption of certainty during the transformation process. This simplification potentially restricts their predictive accuracy and applicability, especially in scenarios characterised by complex and stochastic behaviours of system deterioration and noisy sensor measurements. Additionally, setting deterministic targets for transforming inherently stochastic signals, may significantly increase the complexity and computational time of the applied models (Xu et al., 2023; Chen, Qin, Wang, & Zhou, 2021; Ye, Zhang, Shao, Niu, & Zhao, 2022). These critical deficits motivate our research, highlighting a significant gap in existing methodologies. There is an evident need for enhanced feature extraction techniques that account for the

inherent noise associated with the sensor measurements.
Our research aims to fill this gap by introducing a novel approach integrating a stochastic labelling approach for noisy sensor signals with convolutional neural networks (CNNs). This method leverages Monte Carlo sampling (MCS) of predefined hidden semi-Markov models (HSMMs) to generate high-scoring degradation trajectories that serve as labels for transforming the sensor data utilizing the CNN. By doing so, the model is trained to transform the raw condition monitoring (CM) data into features suitable for prognostic health management (PHM) frameworks while accounting for the inherent noise of sensor measurements. The CNN is also trained on time windows of the data rather than the entire sequences. This attribute enables the model's applicability in real-world scenarios, allowing it to operate online.
The main contribution of the present study is the integration of MCS of HSMMs with CNNs to transform raw sensor data into stochastic degradation trajectories, thus incorporating randomness while being able to operate online in real-world use cases. By alleviating the ill-posed dependency of the FE methods on deterministic labels that stem from intuitive pre-defined metrics, we aim to create a simple and efficient online FE methodology able to transform raw sensor data into features with enhanced prognostic performance. This will ensure the accuracy and higher certainty of the applied PHM frameworks.
The remaining of this study is organised as follows:

- Section 2 delves into the core methodology of this research, providing insight into all of the different components of the proposed transformation methodology as seen in Figure 1.
- Section 3 presents a case study involving acoustic emission (AE) data and demonstrates the proposed methodology's practical application and efficacy. This section is crucial for illustrating the model's ability to handle real-world data.
- Section 4 discusses the research results, by first looking at the transformation of the data and then focusing on the prognostic outcomes. This section highlights the improved prognostic performance achieved using the transformed data by comparing these results against the baseline cumulative transformation of the raw data. This comparative analysis underscores the proposed transformation's enhanced performance in PHM tasks.
- Section 5 concludes the paper with a discussion of the implications of the research findings and proposes ideas for future works.

2. Aspects of the Feature Extraction framework

In this section, we will introduce the method for enhancing the performance of prognostic algorithms by transforming raw data signals into forms similar to those generated through Monte Carlo simulated data. The methodology unfolds over

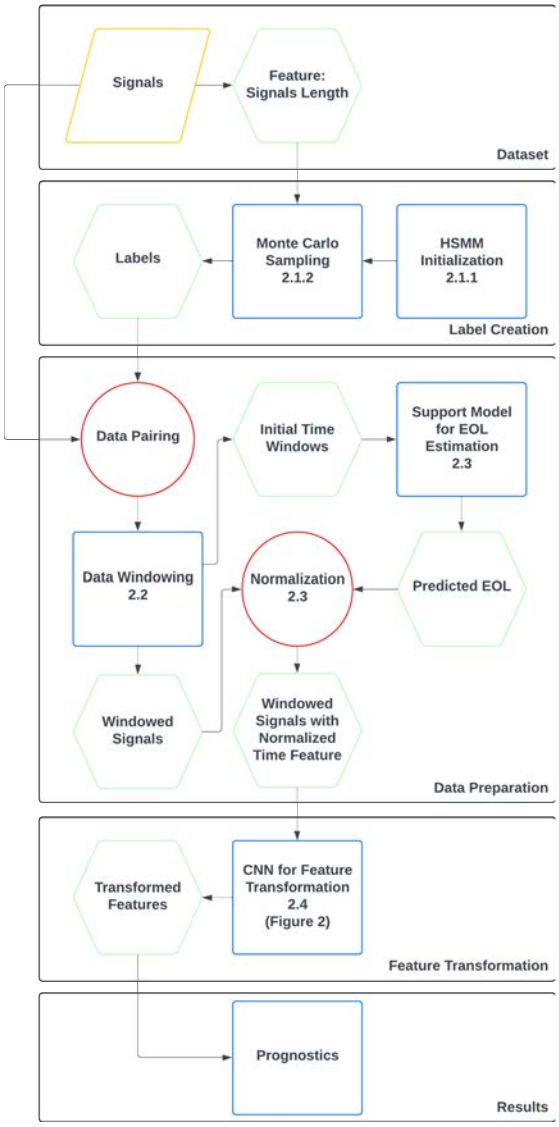


Figure 1. Flowchart of the proposed methodology.

four subchapters, detailing each critical phase of the transformation process:

1. **Label Creation via MCS an HSMM:** This section describes how MCS of HSMM generates labels with the same time length as the target signal. These labels serve as a reference for the desired signal characteristics.
2. **Data Windowing for Online Operation:** The necessity of data windowing is explored here, underlining its importance in developing a system capable of online operation. Data windowing segments the continuous data stream, facilitating real-time processing and analysis.
3. **Support Model for EOL Estimation and time feature normalization:** This part of the method involves creating a model to predict an asset's end of life (EOL). Knowing the EOL is crucial for normalizing the time feature across

all signals, ensuring consistency in the data transformation process.

4. **Convolutional Neural Network (CNN) Model for Feature Transformation:** The technique's culmination is developing a model capable of converting raw data signals into idealized forms. This model leverages the insights gained from the previous steps to enhance the raw signals, making them more suitable for prognostics.

Together, these steps outline a comprehensive approach to data signal transformation that will boost the predictive accuracy and reliability of prognostic tools.

2.1. Label Creation via MCS of HSMM

Our framework's data transformation is based on a supervised learning algorithm. To that end, we need to provide labels for our samples. These labels need to encapsulate the stochastic nature of the assets' degradation. The way to achieve that is by modelling the degradation process with an appropriately initialized HSMM and then utilising it to generate degradation histories of equal length to our training samples. In order to generate degradation histories from a pre-initialized HSMM, we utilize MCS. It is worth noting that the initialization of the HSMM is a short procedure that is explained thoroughly in Sections 2.1.1 and 3.3, and the MCS is a simple algorithm presented in Algorithm 1. Adding to that, the label creation procedure, as explained below, is only required during the training step of the FE procedure. During the online deployment, the transformation of the sensor signals is done in a sub-second time manner since it includes only the windowing of the incoming data and the inference part of the trained CNN. Therefore, the proposed frameworks applicability and scalability is not an issue. In the following, the HSMM's initialisation is discussed, a short introduction to MCS is provided, and finally, its utilization for the required label generation is presented.

2.1.1. HSMM Initialisation

As discussed previously, the first step is to properly initialise the HSMM model to resemble the degradation process. To that end, it's needed first to declare the Initialisation topology ζ (Eleftheroglou, Zarouchas, & Benedictus, 2020):

- **The number of the hidden states (N)**: representing the different levels of degradation.
- **Connectivity between hidden states (Ω)**: This parameter defines the connectivity between the states by defining the allowed transitions between them.
- **Condition Monitoring feature (I)**: The observation of the values of a single CM feature is considered to be the sole indicator of damage in the system.
- **Number of discrete monitoring values (V)**: In the case where the connection between the observation of the CM feature and the damage states is modelled in a non-parametric way, then it has to be converted to several discrete levels V.
- **Transition rate function (λ)**: This is the main characteristic of the degradation process since each transition will follow this function. This parameter can depend on the sojourn time of the current state, the transition between states, the total operating time or any combination of the above

So, in order to fully characterise the HSMM model, a set of parameters $\theta = \{\Gamma, B\}$ are needed where Γ are the degradation process parameters and B are observation process parameters. Γ parameters consist of the parameters needed to define the chosen λ function, and B parameters consist of the emission matrix $\mathbf{B}$. $\mathbf{B}$ is a matrix of dimension of $N * V$ containing the likelihood that every possible observation in the Z space will be emitted by a certain hidden state. After defining the HSMM, in order to generate the required sequences, MCS is utilized, which will be explained in Section 2.1.2.

2.1.2. Monte Carlo Simulated Data

Monte Carlo Sampling is a powerful statistical technique used across various fields. At its core, it leverages the power of randomness to solve complex problems, often too difficult or impossible to tackle with traditional deterministic methods. Monte Carlo Sampling operates on a simple yet profound principle: it uses randomness to approximate problems' solutions. Thanks to the law of large numbers, the more samples are used, the better the actual solution is estimated. Monte Carlo methods are useful when analytical solutions are complex or unavailable, providing a versatile tool for approximation and simulation across diverse applications (Lemieux, 2009). However, Monte Carlo sampling has two main disadvantages. Firstly, it can be computationally inefficient, especially when dealing with high-dimensional or complex problems. Since Monte Carlo methods rely on random sampling to estimate quantities, they may require a large number of samples to achieve accurate results, which can be computationally intensive and time-consuming. Secondly, Monte Carlo methods may struggle to estimate rare events or probabilities accurately with very low or very high values, leading to potential inaccuracies in the results.
In this framework, MCS is utilized to perform a "random walk" over the HSMM. Based on the random sampling and the predefined probability functions of the HSMM, a hidden state is picked for each time step, which in turn emits an observation. The observation is captured, and the "random walk" continues following the design of the model, until the transition to the final observed and terminating state occurs. So, in the context of the proposed methodology, Monte Carlo simulations take place in the training process, so the testing process does not require simulated data. Hence, computational inefficiency is not

a concern for applicability. Additionally, rare samples (events) are not a concern of the feature extraction process since the domain of condition monitoring techniques is predefined in most cases.
Finally, it is worth highlighting that this approach reverses the traditional training process for HSMMs. Instead of relying on multiple observation sequences to estimate parameters, the predefined parameters are used to produce the observation sequences. This method is detrimental to the stochastic generation of trajectories, which are later used as labels for our transformations. By transforming the raw data with these sequences, the predictive accuracy of prognostic algorithms can be significantly enhanced. This is attributed to the fact that this method provides a transformation based on statistical characteristics rather than the traditionally used deterministic labels as explained in the previous. The pseudocode for the implementation of the MCS of the HSMM is adapted from (Eleftheroglou, 2020) and presented in Algorithm 1.

2.2. Data Windowing for Online Operation:

After acquiring the observation sequences that will be used as labels for the transformation of the raw data, our methodology strategically segments both the signals and their corresponding labels into fixed-length time windows. This division is essential for facilitating the model's operation in an online environment, where it is impractical to process the entire signal simultaneously due to the streaming nature of data.

2.3. Support Model for EOL Estimation and time feature normalization

However, the previously mentioned segmentation introduces a challenge: the absence of a definitive feature indicating the end of the signal complicates the transformation process, potentially affecting the accuracy of the model's predictions. To address this issue, we propose developing a secondary model capable of supporting the transformation by predicting the end-of-life (EOL) of the asset, based solely on information from the initial time window. In doing so, we aim to normalise the time feature on a scale from 0 to 1. However, in practice, this normalisation will yield values ranging from 0 to a number close to 1, as it perfomrs only a rough estimation of the EOL, rather than an accurate prediction. This approach aids our model in effectively adapting to and processing signals in real time, paving the way for more accurate and reliable predictions.
To this end, we opted for a fully connected neural network (FCNN) tailored with a specific architecture to meet our predictive objectives. The model is stacked in this order:

- A fully connected layer of 200 neurons is designed to process 200-time-steps inputs of the condition monitoring feature.
- A rectified linear unit (ReLU) function to introduce non-linearity, enhancing its learning capability.
- A dropout layer is then applied to mitigate the risk of overfitting by randomly omitting a subset of neurons during the training phase.
- For the output layer, a single neuron layer is employed to output the predicted time, encapsulating the RUL estimation.

The FCNN is trained only on the first window of the training signals. This model's outputs are then used to normalise the time feature by dividing all time steps of every window by the predicted EOL value. We employ the Mean Squared Error (MSE) as the loss function and Adamax as the optimiser. The training regimen extends over 1000 epochs, with an Early Stopping mechanism in place to monitor progress. This mechanism halts training if no improvement is observed after 50 epochs, simultaneously recovering the best weight combination observed. This approach ensures that the model remains efficient and effective, capturing the essential predictive dynamics without succumbing to overfitting or underfitting tendencies.

2.4. Convolutional Neural Network (CNN) Model for Feature Transformation

After the labels are generated, the signal is split into windows and normalized over its length, the final step is its transformation. This is handled by the primary model, whose architecture is outlined as follows:

- **Convolutional Layer:** This layer has filters, each with a kernel size of 1, facilitating distinct feature detection across time series data without the need for padding. The Glorot uniform method initializes kernel weights, with biases set to zero.
- **Activation Layer:** This layer utilizes the Rectified Linear Unit (ReLU) for non-linear activation, enhancing the model's learning capabilities.
- **Dropout Layer:** A dropout rate is applied in the training phase to reduce overfitting by randomly omitting connections from the previous layer to the next.
- **Fully Connected Layer:** Encapsulated in a time-distributed wrapper, this layer selects the most relevant outputs from the filters, culminating in the final output.

Figure 2 illustrates the architecture of the CNN model, providing an intuitive understanding of its design and flow. This model provides the transformed CM data, concluding the proposed framework.

3. Case Study

In this part, the methodology proposed in Chapter 2 will be applied, showcasing in detail the steps to transform raw acoustic emission data of CFRP specimens under fatigue loading for predicting their RUL. The chapter starts explaining how

Algorithm 1 Pseudocode of Simulated Monte Carlo data generator (Eleftheroglou, 2020)

Inputs:
$M = \{$
ζ (int): model's initialization parameters
θ (array): degradation and the observation parameters
$\}$

Procedure:
$X_0 = 1$
$T_0 = 0$
$T_{age} = 0$
for $(c = 0; c < N; c++)$ **do**
  $i = X_c$
  $s = T_c$
  $j = i + 1$
  $a = U(0, 1)$
  $T_j = \Lambda_{i,j}^{-1}(s, -log(1 - a))$ where $\Lambda_{i,j}(s, t) = \int_0^t \lambda_{i,j}(s, u)\, du$
  $T_{age} = T_{age} + T_j$
  for $(t = T_c + 1; t < T_{age}; t++)$ **do**
    $a = U(0, 1)$
    for $(f = 2; f <= V; f++)$ **do**
      if $\sum_{z=1}^{f-1} b_{X_j}(z) < a < \sum_{z=1}^{f} b_{X_j}(z)$ **then**
        $y_t = f$
      else
        $y_t = 1$
      end if
    end for
  end for
end for

Output:
X_i, T_i (array): respectively the hidden state and the time at the n^{th} transition.
y_t (array): condition monitoring indicator at time $t \in [1, D]$.

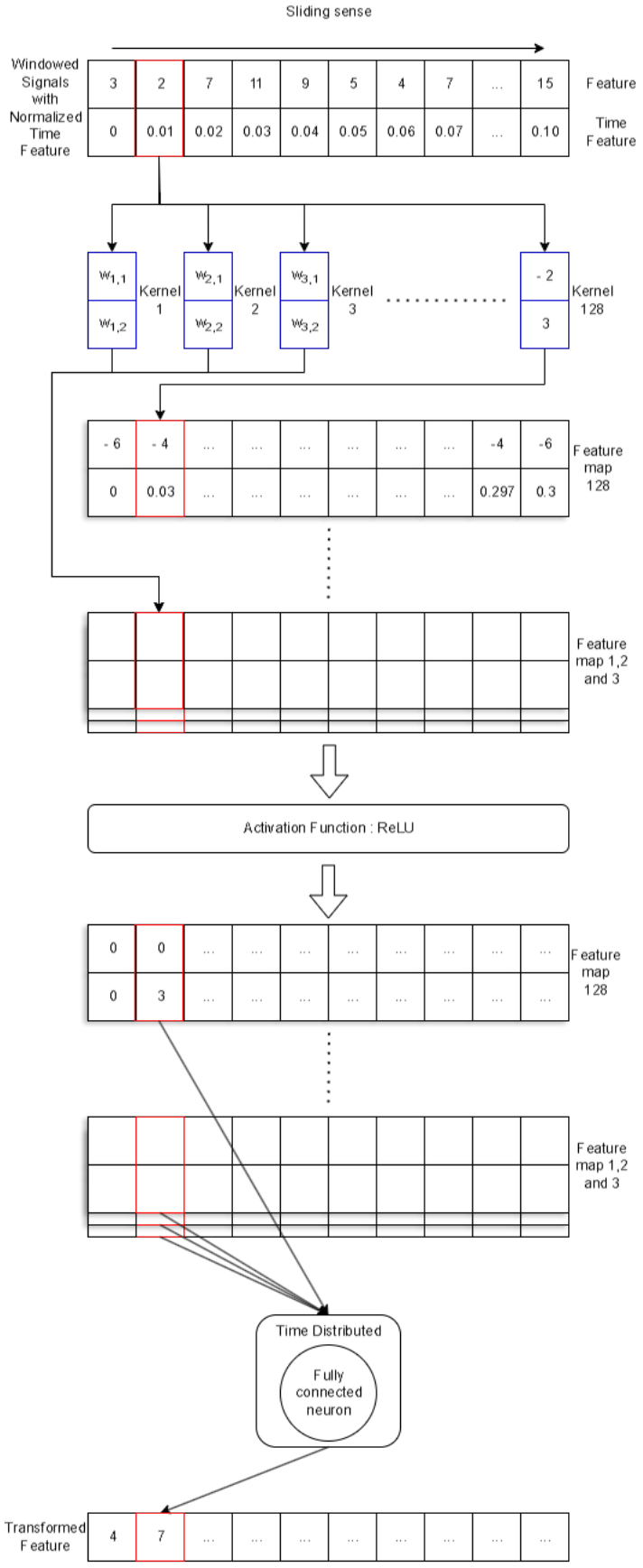


Figure 2. Architecture of the primary CNN model (2.4)

the data are created, preprocessed, labelled, sliced into time windows, and finally transformed.

3.1. Dataset

The acoustic emission (AE) dataset utilised in this research paper comes from the experiments performed in (Eleftheroglou et al., 2020). The dataset is derived from tests on open-hole carbon/epoxy samples, which were subjected to constant amplitude fatigue loading until failure. The specimens of dimensions $400x45mm^2$ are cut from plates manufactured from carbon/epoxy prepregs via the autoclave process. The stacking sequence is a quasi-static lay-up of $[0/45/90/-45]_{2S}$. A hole with a diameter of 10mm was drilled in the center of each specimen. One broadband piezoelectric transducer is attached to each specimen, and with the help of an AMSY-6 Vallen Systeme GmbH, 8-channel AE system, the acoustic emissions of the specimens are recorded and utilized as our CM feature. Hence, the training dataset consists of CM data obtained by utilizing AE sensors from seven samples under fatigue loading, and the testing dataset consists of an eighth specimen that is unseen during training.

3.2. Dataset Preprocessing

The preprocessing phase is designed to enhance the efficiency and effectiveness of model training. The primary objective was to purify the data from noise and scale it appropriately, ensuring that the subsequent steps in our machine-learning pipeline could properly process it. The first required step is to discretize the data. This is achieved by employing the K-means clustering technique (Lloyd, 1982). The entire discretization of the dataset was done by training the K-means with the training dataset, setting the number of clusters to 49, and clustering both the training and test signals with the trained model. This method played a pivotal role in cleaning the data by effectively grouping values into clusters, thereby reducing noise. Each cluster represented a range of values, allowing for a more structured and less noisy dataset. However, the process required careful consideration regarding the number of clusters; an insufficient number could lead to the loss of significant information from the signal. Additionally, this clustering approach ensures that the dataset and the labels are aligned on the same scale. As a final step of the preprocessing phase, we establish a uniform fail value across all signals by setting the final value of each signal feature to 50. The discretized raw data can be seen in Figure 3. By observing the figure, the necessity of transforming the data becomes apparent. The raw feature is highly fluctuating and presents no monotonicity whatsoever. Thus, it cannot be directly used to convey the degradation characteristics of the specimens.

3.3. HSMM initialisation for the case study

This paragraph will discuss the HSMM initialisation required for the Monte Carlo sampling to be performed. To initialise the HSMM, we must define a topology ζ as discussed in chapter 2.1.1.

- **The number of the hidden states (N)** is set at 20 (19 hidden + 1 observed).
- **Connectivity between hidden states (Ω)**: Soft and only left-to-right transitions (meaning that no self-healing or repair actions are modelled) and the final state is observed rather than hidden.
- **Condition Monitoring feature (I)**: The connection between the CM feature's values and the hidden states is modelled with a non-parametric discrete probability function, whose values are defined with the emission matrix in the following.
- **Number of discrete monitoring values (V)** are set at 50.
- **Transition rate function (λ)**: For the model of the degradation of the CFRP specimens under fatigue loading, the Weibull failure rate distribution is chosen as displayed in Equation 1.

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \tag{1}$$

So, in order to fully describe the HSMM model, we need to define the $\theta = \{\Gamma, B\}$ parameters. Γ parameters consist of:

- **Matrix α**: a $(N-1) * N$ matrix of scale parameters for transitions between hidden states. Represented as a diagonal matrix with the first column as zeros. The diagonal is made with logarithmic spacing values from 12 to 6 into 19 values.

$$\alpha = \begin{bmatrix} 0 & 12 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 6 \end{bmatrix}$$

- **Matrix β**: a $(N-1) * N$ matrix of shape parameters for transitions between hidden states. Represented as a diagonal matrix with the first column as zeros as the previous one. The diagonal is made by linear spacing values from 64 to 25 into 19 values.

$$\beta = \begin{bmatrix} 0 & 64 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 25 \end{bmatrix}$$

To complete the initialisation process, we must also create the Emission Matrix **B**. This matrix has dimensions of $N * V$ and displays the probability of the i-th hidden state (rows) emitting

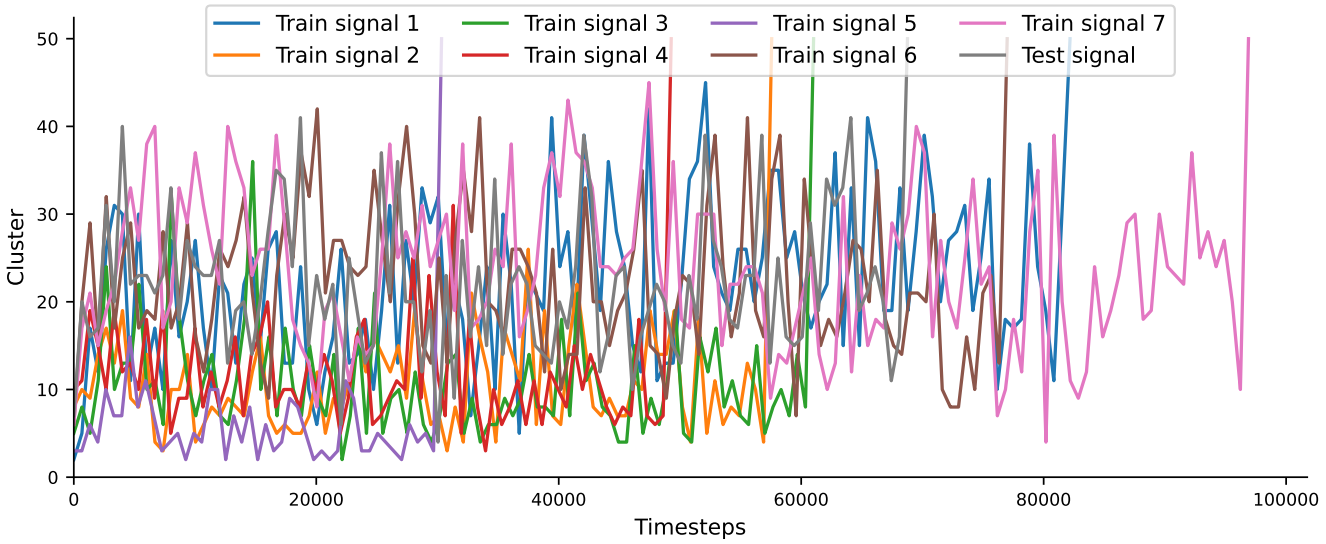


Figure 3. Raw discretized data of the real-world dataset

a specific observable state j-th (columns). The sum of the values in each row always adds up to 1. To fill the values, we have opted for the truncated Gaussian distribution $\mathcal{N}_{T[0,49]}(\mu, \sigma^2)$, which is truncated at 0 and 49 since the failure state is observed, rather than hidden. The standard deviation is equal to 3 for all states, and the mean of $\mathcal{N}_T$ is set in every row in such a way that it increases with the row index. Thus, the mean values of the rows are 19 linearly spaced values in the inclusive $[3, 49]$ range. We have made this decision to ensure that the first hidden states emit the first observable states and the last hidden states emit the last observable states, thus creating a monotonic observation sequence. The emission matrix is presented below:

$$B = \begin{bmatrix} \mathcal{N}^1_{T[0,49]}(3, 3^2) & \mathcal{N}^2_{T[0,49]}(3, 3^2) & \dots & 0 \\ \mathcal{N}^1_{T[0,49]}(5.42, 3^2) & \mathcal{N}^2_{T[0,49]}(5.42, 3^2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

The aforementioned set of parameters is chosen empirically in order for the model to resemble the degradation process of composite specimens under fatigue loading. The user is free to use their own set of parameters that suit their application. So once the parameters are defined, the MCS Algorithm 1 can be run in order to generate the labels.

3.4. Data Windowing of the available data

The data are segmented into fixed-length time windows, as the user specifies. In this research paper, the selected time window is fixed at 200-time steps, corresponding roughly to 10% of the average total length. The size of the windows is chosen intuitively, but it can't be too small since the prediction of the EOL for the normalisation in the following steps will be inaccurate and not too big since doing so will remove the online applicability of the framework. The window advances by one step at each iteration, effectively creating an overlap of 199 timesteps. This allows us to increase the data available for training and testing our transformation model.

3.5. EOL estimations and time feature normalisation of the available data

As previously highlighted, the model's need to predict the EOL derives from our aim to normalize the time feature for the training of the transformation model. The results are presented in Table 1. It can be seen that the error of the estimations varies across the train specimens, with errors up to 20%. However, since this model is used to provide an initial rough estimation of the EOL for the normalization of the time feature, as presented in Section 4, its estimations are more than adequate.

Table 1. EOL Results with Corrected Error Percentages

Signal	Predicted	Actual	Error (%)
Train 1	84128	82176	2.38
Train 2	54304	57568	-5.67
Train 3	48320	61024	-20.82
Train 4	50624	49280	2.73
Train 5	34272	30336	12.97
Train 6	84416	76992	9.64
Train 7	96448	96896	-0.46
Test	77152	68768	12.19

3.6. CNN model architecture for the available data transformation

In developing our 1D CNN model, the choice of hyperparameters, loss function, and optimizer was deliberate and aimed at optimizing performance for our specific dataset characteristics.

- The decision to employ a kernel of dimension 1 ensures that the filter remains unaffected by padding, maintaining the feature map's dimensionality identical to the input. This approach guarantees continuity between successive windows, avoiding noisy spikes at the prediction's beginning and end. Given our transformation goal, this is a critical factor: the prognostic model can lead to wrong predictions.
- For our loss function, Mean Squared Error (MSE) was selected to precisely track the fluctuating nature of our labels, aiming for a regression model that closely mirrors the original data.
- After experimenting with various optimizers, Adamax emerged as the most effective, offering superior convergence properties for our scenario.
- The model architecture was kept minimal with a single CNN layer, a choice driven by the limited size of our signal dataset. This simplicity facilitated a more effective training process compared to deeper models.
- To counteract overfitting due to the high redundancy among the time windows (as detailed in Section 3.4), we implemented L1 regularization and dropout at standard values.
- Due to the non linearity of the labels, CNN is equipped with Rectified Linear Unit.

These choices collectively formed a robust framework for our model, tailored to the unique demands of our data.

3.7. Prognostics

To showcase the effectiveness of the proposed feature extraction for PHM tasks, the Remaining Useful Life (RUL) of the specimens will be predicted by utilising an HSMM. Any prognostic model can be utilized in this step since the data transformation framework is independent of the prognostic model. However, since the degradation process is modelled with an HSMM for the label generation, it is a straightforward choice to utilize a model from the same family. For the prognostics, the explicit duration modification to the HSMM is chosen. Thus, following the initialization procedure explained in Section 2.1.1, the following parameters are defined:

- **The number of the hidden states (N)**: It's considered a hyperparameter of the model, and in order to pick a value, the elbow method utilizing the Bayesian information criterion (BIC) was utilized. The optimal number of states was found to be 7.
- **Transition between hidden states ($\mathbf{\Omega}$)**: soft and left-to-right transitions, no self transitions are allowed.
- **Start probability matrix ($\boldsymbol{\pi}$)**: the process always starts from the first state.
- **Transition rate function ($\boldsymbol{\lambda}$)**: is assumed to be non-parametric and depends only on the current state.
- **Condition Monitoring feature (I)**: The connection between the hidden states and the values of the CM features are assumed to be described by Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$ and represented by a mean and a standard deviation observation vector.
- **CM indicator space (Z)**: since the observation process is modelled with a continuous probability distribution (Gaussian), the indicator space consists of all the real numbers ($Z = \{z \in \Re\}$).

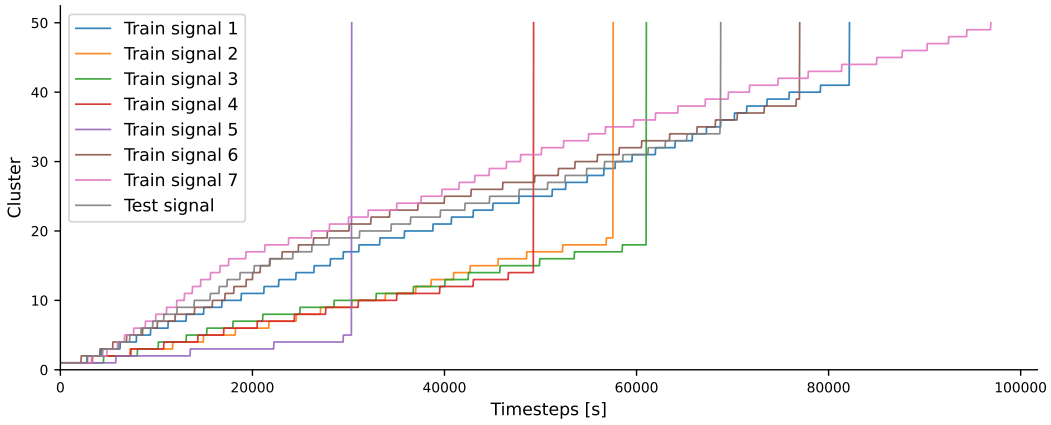
After the parameters are initialised, the parameter estimation can be performed as described in (Yu, 2010). When the optimal parameters have been estimated, they are utilised in order to predict the RUL of the asset, following the procedure in (Dong, He, Banerjee, & Keller, 2006)

4. RESULTS

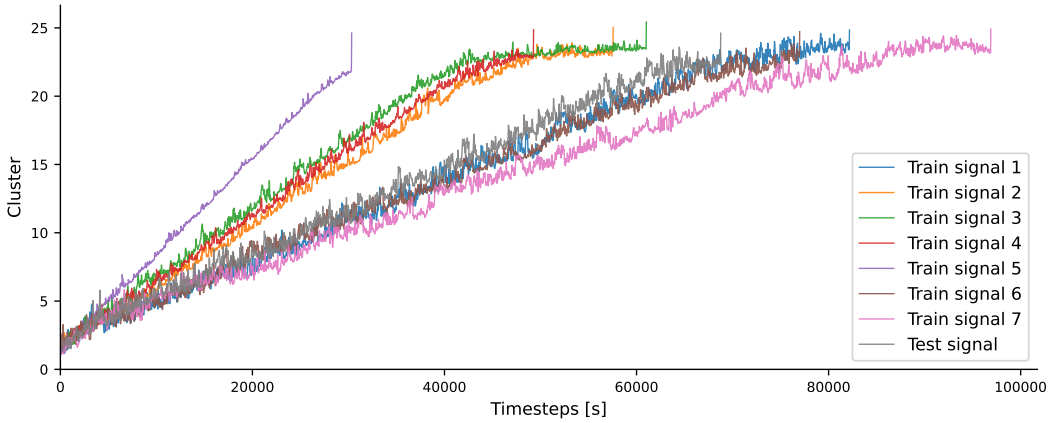
In the first part of the current section, the transformed data utilizing the proposed framework are presented and contrasted against the raw data and their cumulative transformation, highlighting the performance of the framework. Finally, the prognostic findings from the HSMM of the cumulative feature and the one obtained from our framework are contrasted.

4.1. FE results

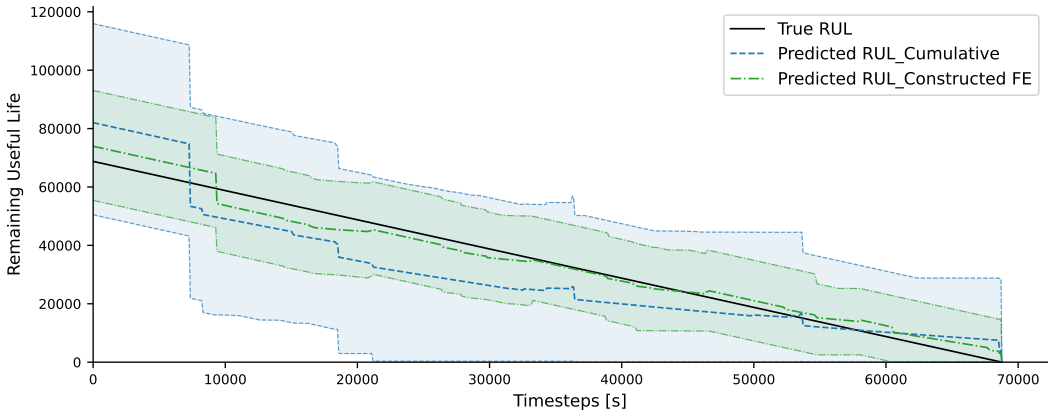
As a baseline for the proposed transformation, the cumulative transformation of the discretized raw data is calculated and presented in Figure 4a. This choice is justified as the authors consider it to be the most straightforward choice for transforming noisy and highly fluctuating data into monotonic ones. The necessity of transforming the data in the first place lies in the inability of prognostic algorithms to provide any meaningful results when applied to fluctuating data that present no monotonic behaviour. In Figure 4b, the transformed data utilizing the proposed framework are showcased. By comparing the two, it can be observed that even though the cumulative features present highly monotonic behaviour (as expected due to the cumulative summation function), there is a great uncertainty associated with the last value directly before failure (also known as prognostability). This is where an added contribution of the proposed transformation also lies. It can be seen (Figure 4b) that the transformed signals (both training and test) fail at values that are very close to each other. This is expected to, in turn, come with reduced uncertainty when it comes to the RUL prediction values, which remains to be seen in the following section.



(a) Cumulative feature



(b) Constructed feature with the proposed methodology



(c) Comparative plot for the test sample

Figure 4. Plots for the results of the proposed methodology

4.2. Prognostic results

In Figure 4c, the prognostic results of the test sample utilising both the cumulative feature and the proposed one are presented on the same plot for comparison reasons. We can see that not only the mean value predictions of the RUL using the constructed feature are closer to the true RUL, but also that the 90% confidence intervals are reduced. This is the manifestation of the main contribution of the constructed feature, which is the reduced uncertainty of the final values of the constructed feature compared to the cumulative one. Hence, the goal of a non-complex FE method that aids in the realization of accurate and highly confident PHM frameworks is achieved.

5. Conclusions

This research introduced a methodology integrating Monte Carlo simulated data with CNN to enhance prognostic performance in predicting system degradation by incorporating the stochastic nature of system deterioration and the noisy measurements in the labels for transforming raw data. Its theoretical viability has been demonstrated, as well as its practical applicability, particularly in its ability to operate online, making RUL predictions for CFRP specimens under fatigue loads, based on noisy AE measurements. It is worth noting that due to the simplicity of the proposed framework, given enough data and a proper HSMM initialization for the MCS (following the procedure showcased in Section 3.3), the applicability to more complex systems is a straightforward procedure. The main contribution of the proposed framework lies in its simplicity of deliberately combining well-established and non-complex components in a novel way that alleviates the deterministic labelling based on intuitively picked metrics in extracting suitable features for PHM applications. This led to the creation of a simple and efficient model that effectively transforms raw and noisy data for accurate and high-confidence prognostics. Our motivation was simple: We consider the labelling of signals that are by nature stochastic with deterministic labels to be ill-posed. Rather, we proposed the generation of labels by sampling a stochastic model (HSMM) in a framework that is independent of the prognostic algorithms and can be applied online. We aim to expand this framework to be able to fuse different CM features and compare its performance against numerous traditional FE methods for prognostic tasks.

References

Chen, D., Qin, Y., Wang, Y., & Zhou, J. (2021). Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing rul prediction. *ISA transactions*, *114*, 44–56.

Coble, J., & Hines, J. W. (2009). Identifying optimal prognostic parameters from data: a genetic algorithms approach. In *Annual conference of the phm society* (Vol. 1).

Dong, M., He, D., Banerjee, P., & Keller, J. (2006). Equipment health diagnosis and prognosis using hidden semi-markov models. *The International Journal of Advanced Manufacturing Technology*, *30*, 738–749.

Eleftheroglou, N. (2020). *Adaptive prognostics for remaining useful life of composite structures* . Delft University of Technology. (Thesis (Ph.D.))

Eleftheroglou, N., Zarouchas, D., & Benedictus, R. (2020). An adaptive probabilistic data-driven methodology for prognosis of the fatigue life of composite structures. *Composite Structures*, *245*, 112386.

Lemieux, C. (Ed.). (2009). *Monte carlo and quasi-monte carlo sampling*. Springer.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Moradi, M., Broer, A., Chiachío, J., Benedictus, R., & Zarouchas, D. (2023). Intelligent health indicators based on semi-supervised learning utilizing acoustic emission data. In P. Rizzo & A. Milazzo (Eds.), *European workshop on structural health monitoring* (pp. 419–428). Cham: Springer International Publishing.

Xu, Z., Bashir, M., Liu, Q., Miao, Z., Wang, X., Wang, J., & Ekere, N. (2023). A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model. *Computers & Industrial Engineering*, *176*, 108999.

Ye, Z., Zhang, Q., Shao, S., Niu, T., & Zhao, Y. (2022). Rolling bearing health indicator extraction and rul prediction based on multi-scale convolutional autoencoder. *Applied Sciences*, *12*(11), 5747.

Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial intelligence*, *174*(2), 215–243.

Development of a PHM system for electrically actuated brakes of a small-passenger aircraft

Riccardo Achille[1],Andrea De Martin[2],Antonio Carlo Bertolino[3], Giovanni Jacazio[4], and Massimo Sorli[5]

[1,2,3,4,5]*Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Torino, 10129, Italy*

riccardo.achille@polito.it
andrea.demartin@polito.i
antonio.bertolino@polito.it
giovanni.jacazio@formerfaculty.polito.it
massimo.sorli@polito.it

ABSTRACT

The evolution towards "more electric" aircraft has seen a decisive push in the last decade, due to the growing environmental concerns and the development of new market segments (Urban Air Mobility). Such push interested both the propulsion components and the aircraft systems, with the latter seeing a progressive trend in replacing the traditional solutions based on hydraulic power with electrical or electro-mechanical devices. Electro-mechanical brakes, or E-Brakes hereby onwards, would present several advantages over their hydraulic counterparts, mainly related to the avoidance of leakage issues and the simplification of the system architecture. Moreover, although it is expected a weight increase of the brake, the elimination of the hydraulic lanes would still come with an overall weight reduction. Despite these advantages, it remains a new, relatively unproven technology within the civil aviation field. Within this context, the development of PHM solutions would align with the need for an on-line monitoring of a relatively unproven component. This paper deals with the preliminary stages of the development of such PHM system for the E-Brake of a future executive class aircraft, iterating on previously published material and presenting a particle filtering approach based on a new degradation model and data provided through a revised high-fidelity model. The paper opens with the introduction to the research project and the technological demonstrator, positioning the performed work within the available literature. PHM activities, performed on simulated data-set are then presented and the preliminary results discussed.

Keywords: PHM, EMAs, Brakes, Particle filter.

1. INTRODUCTION

Electro-mechanical brakes, or E-Brakes, are the next step in the evolution of aeronautical braking systems, and the natural consequence of the push for the electrification of civil aviation which is strongly affecting the development of future platforms. E-Brakes have already found applications in civil aviation, with different architectures already flying on the latest iterations of the Boeing 787 and the Airbus A-220 and have drawn interest in the aeronautic community thanks to the significant advantages over their hydraulic counterpart, including lower weight, the elimination of long hydraulic pipelines and a lower environmental imprint. Despite these success stories, they remain a relatively unproven and more complex technology with respect to the traditional hydraulic solution. The definition of a comprehensive PHM system, leveraging the higher number of sensors usually employed on electro-mechanical system, would provide additional confidence towards their application, lowering the risk of unanticipated failures, reducing the aircraft downtimes and giving access to strategic information useful to optimize the fleet management. Although literature on PHM activities for the most common components of electro-mechanical brakes is extensive, few papers have been published about the E-Brakes themselves. In (Ramesh et al., 2021) authors propose a Fault Detection and Identification (FDI) algorithm to observe and correctly assess the most probable failures occurring in a simple electro-mechanical brake for aeronautic applications. The analysis considers an aeronautical brake actuated by one Electro-mechanical actuator driven through a brushed DC-motor and is mainly focused on electrical failures. In (Oikonomou et al., 2022) authors investigate the prognosis of wear in aeronautical brakes through the analysis of historical series of brake pads thickness. Data-driven techniques are applied to perform the long-term prognosis, and the results of an interesting benchmarking activities comparing the performances of several algorithms are provided. Results are promising but assume the presence of

dedicated sensors to measure the thickness of the brake pads, which are not foreseen for the application under study in this paper. The E-LISA research project, performed within the Clean Sky 2/Clean Aviation framework, has the objective of developing an innovative iron bird dedicated to executing tests on the landing gear of a small aircraft equipped with an electro-mechanical landing gear and electrical brake. The E-LISA iron bird consists of a multi-functional intelligent test facility integrating hardware and software, allowing all the tests and analyses perceived as fundamental to be performed to demonstrate the maturity of an electro-mechanical landing gear, hence paving the way for its implementation in a small passenger aircraft. Such tests include the simulation of complete landing procedures under different operating conditions such as runway friction (wet/dry), presence of waving and irregularities along the runway, variable aircraft weight, and approach speed. At the same time, the rig will act as a technological demonstrator for PHM routines devoted to the analysis of the E-Brake health status. This paper opens with the description of the case study under analysis, a fully electrical landing gear leg for a new executive-class aircraft, detailing the system characteristics. Then the architecture of the technological demonstrator is presented, highlighting its most prominent features and the solutions required to meet its functional requirements. The focus is then shifted towards the definition of the PHM routines, where a possible scheme to detect and prognose the wear of the brake pads is proposed. A first-tentative approach to the problem was presented by authors in (De Martin, Jacazio, Parisi, et al., 2022), making use of a first-trial high-fidelity model and a simplified particle filtering approach. This study was used a basis: the model was further evolved to take into consideration the effects of different combination of environmental conditions and tires types, while the prognostic algorithm was substantially modified considering a physics-based representation of the fault progression. Early results are presented, alongside a test-plan to support the simulation findings through experimental activities, once the rig is operational.

2. CASE STUDY

The case study under analysis is an E-Brake system for an executive-class aircraft with an expected weight at take-off ranging between 5.5 and 6.1 tons, depending on the passengers' number and the amount of unspent fuel. Two E-Brake systems are integral with the Main Landing Gear system, one for the Left-Hand side, one for the Right-Side each. Depicted in Figure 1, each E-Brake is a multi-disk assembly actuated through four Electro-Mechanical Actuators (EMAs) controlled in force. Whenever the pilot acts on the brake pedals, a force command is sent towards the E-Brake system; such command signal is processed by the Brakes Control Unit (BCU), which can cut the force command signal through the touch-down protection routines, avoiding that the brakes are actuated before the aircraft rotation during landing has ended. The command signal can

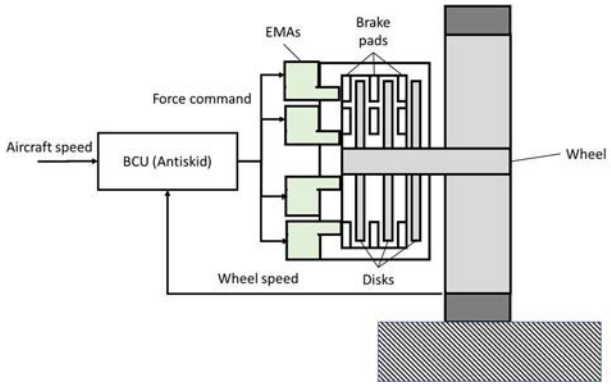


Figure 1. Case-study architecture.

be further modulated by the electronic anti-skid system, which decreases the force request depending on the runway conditions to avoid the occurrence of wheel blockage events and excessive slip according to a combination of pilot input and automatic recognition of the runway status. The electro-mechanical actuators are driven by one Brushless-DC motor with each, and act on the brake pads through a mechanical transmission made of a one-stage reducer and a ball-screw. Each actuator is equipped with a force sensor to measure the exerted action, while a resolver integral with the motor shaft is used to infer its position and realize the Field Oriented Control of its phase currents. The test-rig has being design to exchange information with the E-Brake during its operations interfacing with the BCU. As such all the signals provided by the sensors employed in each EMAs are acquired and can be used for PHM. Such signals include the angular position of the E-Brake motors shaft, the braking command issued by the pilot, the anti-skid and touch-down protection signals and the exerted force measurement for each EMA. The phase currents of each electric motor are acquired as well, along with the current request provided by the force control loops.

3. THE TECHNOLOGICAL DEMONSTRATOR

The main purposes of the technological demonstrator are to support the testing and certification of a novel E-Brake system and to foster the definition of dedicated prognostic logic. As such, the main functional requirements are twofold. On one side it is paramount to be able to conduct the tests prescribed by the normative, such as the on-ground start-stop test and the landing procedure. On the other side, it is important to recreate on the test bench the widest array of operating conditions to stress as much as possible the PHM routines and provide statistically representative datasets. To achieve this goal the test rig is designed to allow the arbitrary variation of several operating parameters, including the aircraft approach speed, the dynamic load applied on the landing gear leg, the friction coefficient between the aircraft wheels and the runway and the occurrence of a selected number of electrical failure or mechanical faults. The architecture of the test rig is depicted in Figure 2, while a

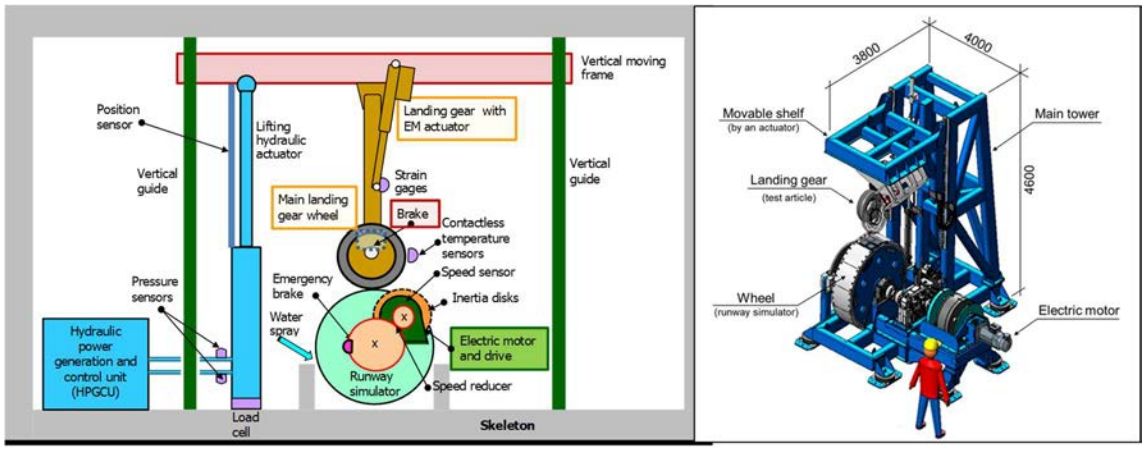


Figure 2. Iron bird schematics.

details of its mechanical structure can be retrieved in (Giannella et al., 2022). The mechanical structure can be divided between a fixed part and a moving platform integral with the landing gear leg, complete with a wheel and electrical brake. The moving platform can translate vertically along low friction guides according to the force provided by an electro-hydraulic servoactuator controlled through a 160 l/min servovalve. A calibrated by-pass orifice connects the two hydraulic lines serving the actuator to improve the dynamic response of the force-controlled system. The test-rig behavior is continuously monitored through one linear variable differential transformer (LVDT) sensor measuring the hydraulic actuator travel, a load cell measuring the force exchanged between the actuator and the moving platform and a differential pressure transducer sensing the pressure drop across the two actuator's chambers. The hydraulic power available for the test-rig operation is that of the facilities in which the rig will be installed and is limited at 207 bar. The contact between the landing gear wheel and the runway is represented through a runway simulator, a rotating disk, connected to a selected number of inertia disks, representative of the aircraft inertia, through a gearbox. A different solution, based on a novel hydraulic system, was considered in (De Martin, Jacazio, Ruffinatto, et al., 2022) but discarded due to budget constraints. The diameter of the rotating cylinder is such to be representative of the expected linear speed of the aircraft along the runway during the landing procedure and must be higher than the wheel diameter to reduce at a minimum the differences between the wheel/runway simulator contact and the wheel/real runway contact. A gearbox is interposed to significantly reduce the mass and the encumbrance of the flywheels, the number of which can be increased or decreased to scale-up or scale-down the weight of the simulated aircraft. The runway simulator was designed with the possibility to change the external coating. To achieve the variation of the friction forces between the wheel and the runway and allow the verification of the anti-skid logic behavior in different operating conditions, while a sprinkler can be activated to reproduce the wet-runway conditions. An electric motor is used to accelerate the runway simulator up to the angular frequency corresponding to the aircraft horizontal speed given the diameter of the rotating disk, while an emergency brake is installed in-line with the rotating cylinder, allowing to bring the full system to a complete stop in less than 60 s. The technological demonstrator is controlled through an engineering test station (ETS), which accepts the inputs from a central control unit (CCU) that in turn receives the commands from an operator via a user interface. The input signals are then sent together with rig measurements to a dedicated computer running a real-time (RT) representation of the aircraft dynamics during landing. Such real-time model is then used, along with a model of the runway and a model of the landing gear dynamics, to compute in real-time the load that must be applied to the test-article. The ETS also include the rig control logic, which is designed to manage both the position of the moving platform and the force exerted by the hydraulic actuator and include a safety routine check to limit the damages to the rig and to the test article in case of a failure of the anti-skid system during the execution of a test. The structure of the control system is depicted in Figure 3, where three main modules can be identified. The Simulation Module involves the real-time representation of the landing dynamics, including a real-time representation of the aircraft dynamics, a runway model, which allows to describe the presence of periodical or localized runway irregularities, and the model of the landing gear legs, each modelled as a two-degrees of freedom vibrating systems, where the mechanical

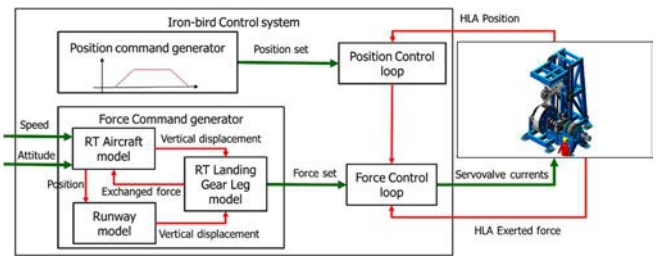


Figure 3. Control system structure.

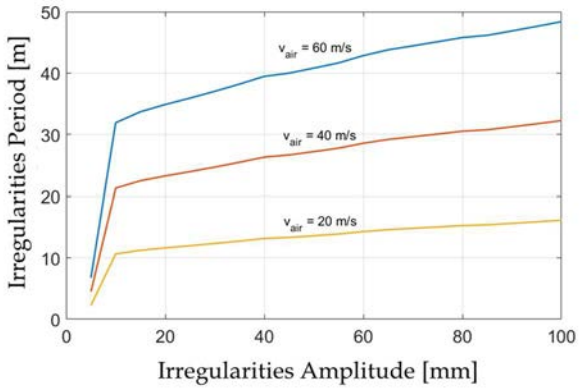


Figure 4. Projected limits on the representation of the periodical runway irregularities on the test-rig.

characteristics of the shock-absorber, of the tires and of the mechanical structure are provided by the industrial partners of the project. The control system is based on two control loops, one operating in position and the other on force, alternatively active before and after the contact between the landing gear wheel and the runway simulator. The control system, best described in (Bertolino et al., 2023) is designed to achieve bandwidth higher than 10 Hz in both force and position control loops. As shown in Figure 4, the control system, combined with the Real-Time simulation of the aircraft behavior during landing, it allows to reproduce on the test-rig the effects of periodical runway irregularities up to 100 mm depending on their period, the aircraft mass and its expected approach speed (De Martin, Jacazio, & Sorli, 2022).

4. Simulation Activities

To achieve the definition of a PHM scheme for the entire E-Brake assembly a number of tasks are required. In order, there is the need to orderly assess which failure modes or which cause of service disruption to investigate, to prepare a high-fidelity model of the system representative of its operational performances and finally proceed to support such activities through experimental tests and validation. As stated in the introduction the technological demonstrator is not yet in function, thus the analysis is so far limited to simulation results. The failure modes being investigated includes the occurrence of several types of short circuits within the windings of the electro-mechanical actuators and wear of the mechanical transmission. Such failure modes, although important, are fairly common in electro-mechanical actuators independently from their function. As such, priority was given to the detection and prognosis of the wear of the brake pads. Although not a failure-mode per se, the reduction of the brake pad thickness is accompanied with a lower dynamic response of the brake and requires periodical maintenance operations. Currently wear of the brake pads is detected through dedicated sensors or through periodical inspection of visual indicator on top of the brake itself. Making the monitor of the brake pads wear part of the PHM system would allow to limit or avoid the necessity of periodical inspections, anticipate the maintenance action and avoid unpredicted aircraft-on-ground situations. Authors started analyzing the possibility of such a system in (De Martin, Jacazio, Parisi, et al., 2022) through a simple particle filtering routine based on data coming from streamlined landing simulations. As more data were made available the analysis was improved through the definition of a higher-fidelity simulation model and a more realistic operational scenario. The prognostic routine was similarly revised and will be presented in the next sections of the paper.

4.1. System modelling

The high-fidelity simulation model includes a two-dimensional representation of the aircraft dynamics during landing, a three degrees-of-freedom multi-body model of the landing gear legs and of their interaction with the runway, and an in-detail representation of the E-Brake dynamics. Parts of such model, with particular reference to the aircraft dynamics and the vertical displacement of the landing gear leg were already described in (De Martin, Jacazio, & Sorli, 2022; De Martin, Jacazio, Parisi, et al., 2022) and won't be reported here, limiting the description to the component interested by modifications. Starting with the rotational dynamics of the wheel and addressing with $F_n = k_t(x_w - x_{rw}) + c_t(\dot{x}_w - \dot{x}_{rw})$ the vertical force exchanged between the wheel and the runway it is possible to express the wheel angular acceleration $\ddot{\vartheta}_w$ as a function of the rolling friction coefficient u_{rw}, expressed as a function of the wheel angular frequency and of the tire pressure (Carbone & Putignano, 2013), of the the moment of inertia of the wheel assembly I_w, the wheel diameter D_w and the viscous friction coefficient roughly representative of the dissipation in the wheel supports c_w.

$$F_n\mu\left[\frac{D_w}{2} - \left(x_{leg} - x_{rw}\right)\right]\text{sign}(\lambda) - F_n u_{rw}\tanh\dot{\vartheta}_w - c_w\dot{\vartheta}_w - T_{brk} = I_w\ddot{\vartheta}_w \quad (1)$$

The friction coefficient μ is evaluated according to a modified version of the Burckhardt model (M. Burckhardt, 1993), as a function of the slip factor λ between wheel and runway simulator and of the experimental parameters β_1, β_2 function of the tires temperature, thread type and runway conditions.

$$\mu(\lambda) = \left[\beta_1\left(1 - e^{-\beta_2\lambda}\right) - \beta_3\lambda\right]\mu_{max}(p_{tire}, v_{air}) \quad (2)$$

The parameter $\mu_{max}(p_{tire}, v_{air})$ is function of the tires pressure and of the aircraft speed and was fitted on experimental dataset provided for different combination of thread type and runway conditions for an aircraft of similar size to the target platform. An example of the fitter profiles is shown in Figure 5, where data for smooth tires on dry-runway

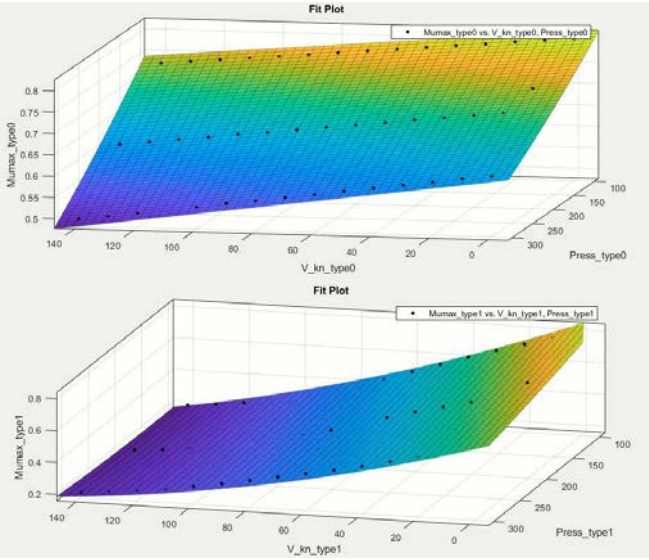


Figure 5. μ_{max} values for dry (above) and wet (below) runway conditions.

conditions are compared against data for threaded tires on a wet runway. The four Electro-Mechanical Actuators (EMAs) responsible for the braking action are controlled in force and act in parallel on the multi-disk brake. The control system is described as a two-nested control loops, where a sequence of Proportional-Integrative controllers operates on the force control loop and on the current control loop of each brushless motor. The sensors are modelled through second order transfer functions replicating the expected dynamics of the load cell and of the Hall-effect sensors employed to monitor the angular position of the Brushless-DC rotor. The simulation of the measure chain is complete with the model of the employed A/D converters. The dynamic model of each EMA features a functional description of the Electronic Power Converter derived from (Mohan et al., 2005) for a three-phase inverter controlled through Pulse Width Modulation (PWM). The electrical dynamics of the motor is described according to a streamlined three-phase model of the system, where $V_{a,b,c}$ and $i_{a,b,c}$ are the phase voltages and currents.

$$[V_{a,b,c}] = [\boldsymbol{R_{a,b,c}}(T_w)][i_{a,b,c}] + [\boldsymbol{L}(T_w)]\frac{d}{dt}[i_{a,b,c}] + \frac{d}{dt}[\phi_{a,b,c}(\vartheta_{el})] \quad (3)$$

$[\mathbf{R_{a,b,c}}]$ is the electric resistance matrix, which elements depends on the windings' temperature (T_w). The diagonal elements of the matrix represent the single-phase resistance, while the non-diagonal terms represent the electrical resistance provided by the insulating material separating the coils of different phases, addressed as phase-to-phase resistances hereafter. Such values can be degraded to simulate the effects of a turn-to-turn or phase-to-phase short respectively. [**L**] is the inductance matrix, accounting for self-induction and mutual induction phenomena along with the effect of magnetic flux dispersion. Finally, $[\phi_{a,b,c}]$ is the concatenated magnetic flux provided by the permanent magnets, function of the electrical angle (ϑ_{el}). The torque at the motor shaft can then be computed, leading to the dynamic equilibrium of the rotor

$$\sum_{a,b,c}\frac{d\phi}{dt}i_{a,b,c} - c\dot{\vartheta}_m - k_m(\vartheta_m - \vartheta_{gb}) - c_m(\dot{\vartheta}_m - \dot{\vartheta}_{gb}) = I_m\ddot{\vartheta}_m \quad (4)$$

where ϑ_m and ϑ_{gb} are the angular position of the motor shaft and of the gears. I_m is the moment of inertia of the rotor, while k_m and c_m address the torsional stiffness of the motor shaft and its associated damping. The gear pair is described as a rotational mass-spring-damper system, thus leading to the following equation,

$$k_m(\vartheta_m - \vartheta_{gb}) + c_m(\dot{\vartheta}_m - \dot{\vartheta}_{gb}) - \frac{1}{\tau}[k_{gb}(\vartheta_{gb} - \vartheta_{rs}) + c_{gb}(\dot{\vartheta}_{gb} - \dot{\vartheta}_{rs})] - T_{fr,gb} = I_{gb}\ddot{\vartheta}_{gb} \quad (5)$$

where τ is the transmission ratio, $T_{fr,gb}$ the friction torque, while ϑ_{rs} is the angular position of the rotating part of the screw. The friction torque is computed as the sum of three components, one dependent on the acting load, one related to the viscous friction and a drag torque component. The power-screw is modelled as a two-degrees of freedom elements, where the rotating part is connected to the translating element through a viscoelastic element. Defining with $x_{rs,i}$ the position of the translating portion of the screw pertaining to the i-th actuator, it becomes possible to describe the brake dynamics, and thus that of the pads. Addressing with k_{eb} the stiffness, it is possible to evaluate the braking torque acting on the landing gear wheel as a function of the translating mass of the brake pads m_{eb}, its translation x_{eb} and the angular speed of the wheel $\dot{\vartheta}_w$ as,

$$\begin{cases} T_{brk} = 0 \leftrightarrow x_{eb} < x_{thr} \\ T_{brk} = R_{eb}f_{eb}[k_{eb}(x_{eb} - x_{thr}) - c_{eb}(\dot{x}_{eb})]x_{eb} \geq x_{thr} \end{cases} \quad (6)$$

where $f_{eb} = f_{eb}(\dot{\vartheta}_w)$ is the friction coefficient between the brake pads and disk, function of the wheel angular frequency. Knowing the braking torque and the wheel angular frequency it is possible to compute the mechanical power transformed into heat by the braking process. Such power is used within a simplified thermal model of the E-Brake assembly to estimate at each time step the temperature of the pads and the temperature of the electric motor windings considering both the thermal power generated by the motor themselves and that transmitted to the external environment. Since the pads contact the brake disks only when their translation x_{eb} overcomes a predefined stroke equal to x_{thr}, it is possible to model the effects of the pads wear by properly increasing such threshold value under the assumption that the brake pads return in the original position once the braking procedure is

finished. According to (Olesiak et al., 1997; Yevtushenko et al., 2017) , wear progression in brake pads can be described as dependent on an experimental coefficient f_{wear} and k_{wear}, function of the local absolute temperature T, the sliding velocity between disks and pads v, and the contact pressure p. The dependency of the friction and wear coefficient on temperature is due to how the interaction between the disc and the brake pads occurs at a microscopic level. The brake performance (and its wear) depends on how the material of the pads bonds with the material of the disk and with the particles of the pads material which have remained bonded with the disc due to the run-in process and previous usage. This process is temperature dependent: at very-low temperature (-40°C) friction tends to increase in mechanical systems. Similar effects occurs at very high temperatures, where the bonding effects can be favored by localized fusion processes, accelerating the wear rate.

$$\Delta x_{thr} = \int_t f_{wear}(T) K_{wear}(T) v(t) p(t) dt \tag{7}$$

Expressing the sliding velocity as a function of the wheel angular frequency $\dot{\vartheta}_w$ and the radial coordinate of the pads with respect to the wheel axis R_{pad}, we have

$$v = \dot{\vartheta}_w R_{pad} \tag{8}$$

The average pressure within the pads/disks contact area can be computed as a function of the braking force exerted by the four actuators and the pad contact area.

$$p = \frac{k_{eb}(x_{eb} - x_{thr}) - c_{eb}(\dot{x}_{eb})}{A_{pad}} \tag{9}$$

An example of the model response is provided in Figure 6, where the system behavior in response to an emergency brake in presence of a wet runway and smooth tires is presented.

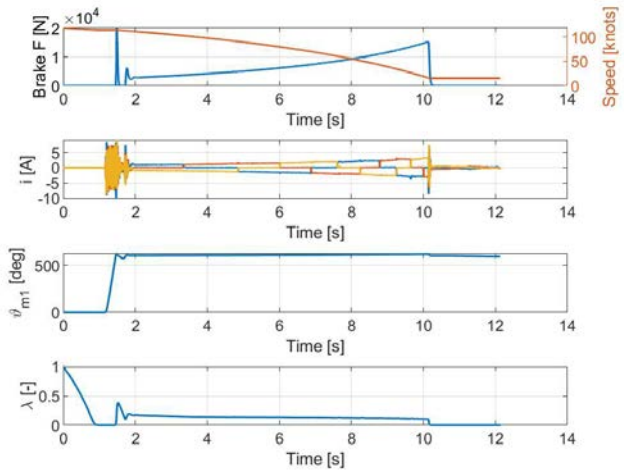


Figure 6. Example of the model response.

The figures depict the behavior of the total braking force expressed by the four EMAs, the aircraft speed along the runway, the three phase currents of the BLDC motor belonging to one of the EMAs (i [A]), the correspondent angular rotation of the motor shaft ϑ_{m1} and the slip factor λ between the wheel and the runway.

4.2. Operational scenario and data-base building

The simulation model is built for two main reasons: to generate a data-base to train and stress the PHM routines and to help in characterizing the system and provide additional information that could be useful to the PHM system itself. The first task in generating a reliable operational scenario is to characterize all possible sources of uncertainty of the system behavior. For the case study under analysis the following sources were identified and addressed.

- Aircraft mass at landing
- Runway temperature and conditions
- Tires type (smooth, threaded) and pressure
- Aircraft horizontal approach speed
- Type of braking procedure (emergency, normal)
- Production tolerances in the E-Brake system
- Pilot reaction time
- Sensors noise, deviations

The aircraft mass at landing is drawn randomly at each simulation from a uniform distribution ranging between 5.5 and 6.1 tons approximately, which is considered the expected variance depending on the passengers number, payload presence and type, and the remaining fuel. Runway temperature and conditions were taken into account by considering the temperature and rainfall distribution of three distinct area, each representative of a prevalently cold (Vancouver), hot (Dubai) and temperate (Rome) operating conditions. Data were obtained through public access data-base and randomly drawn at each simulation. Tires type is chosen at each simulation between threaded and smooth, while their pressure is randomly chosen from a normal distribution with mean 200 psi and variance 30 psi. The aircraft horizontal approach speed is randomly drawn from a normal distribution with mean equal to 110 knots and standard distribution equal to 5 knots. The type of braking procedure (whether emergency or normal) is decided before launching the simulation: during emergency stops, the pilot commands the E-Brake to supply its maximum force, while the anti-skid routine modulates such request to achieve the slip factor associated with the maximum friction coefficient between wheel and runway. On the contrary, during normal operations, the pilot modulates the braking force command trying to mimic a sequence of aircraft horizontal speed drawn from a pool of experimental results provided by the industrial partners of the project for an aircraft of similar size. An example of the difference between the two braking procedures is provided in Figure 7.

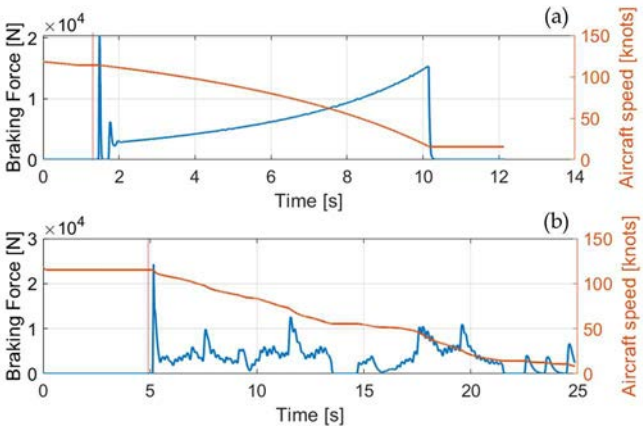


Figure 7. Comparison between emergency stop (a) and normal braking (b).

In this comparison, both aircrafts are decelerated down to 16 knots, which is the limit taxing speed on most civil runways. Another uncertainty source related to the brakes performances is the pilot reaction time, which is modelled as a simple transport delay with time constant variable between 0.1 and 0.5 s.

These first uncertainty sources are expected to be the most significant for the E-Brake performances.

The aircraft speed and its mass at landing affects the overall kinetic energy to be dissipated through the brakes, while the operating conditions, tires type and inflating pressure affects the efficiency with which the braking torque produced by the electro-mechanical device is transferred to the ground.

Additional uncertainty sources, affecting the signals pertaining the E-Brake that can potentially be used for PHM includes the production tolerances, which were over imposed onto the main electrical and mechanical parameters of the EMAs, and sensors noise, modelled according to the indication provided in the manufacturer catalogues.

5. PHM ALGORITHM

The PHM algorithm for the wear of the brake pads is based upon the scheme presented in Figure 8, designed to leverage the physical knowledge of the system and employ the past knowledge of the E-Brake operations to better characterize the uncertainty distribution.

Prognosis is achieved through a Bayesian estimation method using a particle filtering approach, which provides after each landing an estimate of the current level of wear in the brake pads leveraging the indirect knowledge of the terms employed in the wear function provided in Equation 7. This step, functionally part of the fault identification process, is necessary to achieve the long term prognosis through physics-driven equations by supplying the particle filter with pdfs of future usage of the brake based on previous information retrieved and stored after each landing. Particle filters, firstly introduced in PHM by (Orchard & Vachtsevanos, 2009), take advantage of a nonlinear process (fault / degradation) model to describe the expected dynamics of the fault progression and a measure model derived from the feature/wear progression dependence observed during the feature selection phase. The particle filtering approach was chosen over other considered options (LSTMs in particular, were considered following previous works on a different application (Grosso et al., 2020)) for several reasons. The main one is that this process enables the estimation of the fault size through physics-based equations; since activities have been performed over simulated data-set, the adoption of a physics-driven approach was preferred to mitigate the risk of not operating on experimental datasets. The second reason is that brake pads wear is only one of the many failure modes that can occur in electro-mechanical brakes. Other failure modes, such as the brake rotors wear, faults in the EMAs mechanical and electrical components or issues in the EMAs sensors can occur, and have a direct effect on the progression of other failure modes. Looking ahead, the authors have planned to pursue the definition of a PHM scheme able to

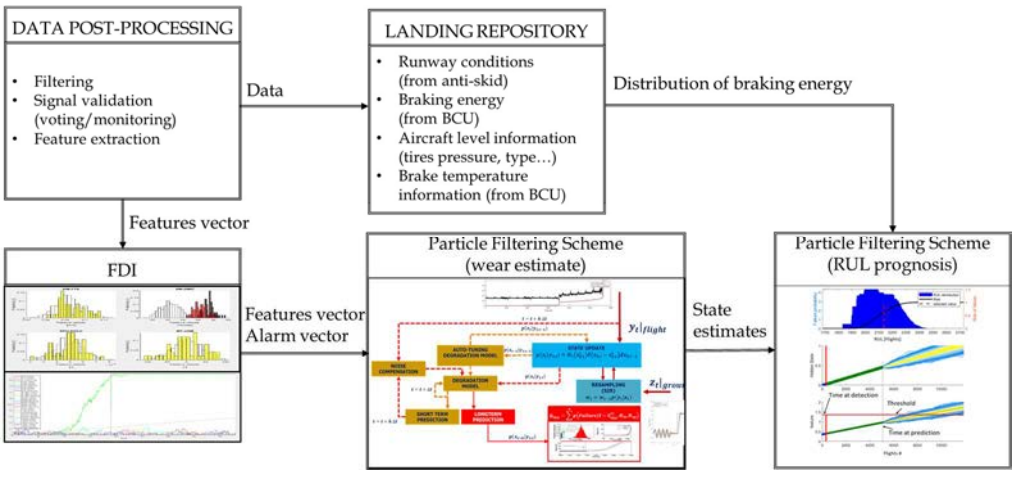


Figure 8. PHM Scheme.

define the expected RUL of the system including the possible cross-dependence between different failure modes. Particle-filtering techniques naturally lend themselves to this step, since their results are well-suited to data-fusion algorithms (Vachtsevanos et al., 2006). Moreover, being based on a physics-based description of the failure mode, the outcome of PF algorithms can be more easily interpreted from an engineering perspective and more easily included in digital-twin representations of the system under analysis.

Prognosis through particle filtering is achieved by performing two sequential steps, prediction and filtering. Prediction uses both the knowledge of the previous state estimate and the process model to generate the a priori estimate of the state probability density functions (pdfs) for the next time instant,

$$p(x_{0:t}|y_{1:t-1}) = \int p(x_t|y_{t-1})p(x_{0:t-1}|y_{1:t-1})\,dx_{0:t-1} \quad (10)$$

This expression usually does not have an analytical solution, requiring Sequential Monte Carlo algorithms to be solved in real-time with efficient sampling strategies (Roemer et al., 2011). Particle filtering approximates the state pdf using samples or "particles" having associated discrete probability masses (often called "weights") as,

$$p(x_t|y_{1:t}) \approx \tilde{w}_t(x_{0:t}^i)\delta(x_{0:t} - x_{0:t}^i)dx_{0:t-1} \quad (11)$$

where $x_{0:t}^i$ is the state trajectory and $y_{1:t}$ are the measurements up to time t. The simplest implementation of this algorithm, the Sequential Importance Re-sampling (SIR) particle filter (Arulampalam et al., 2009), updates the weights using the likelihood of y_t as:

$$w_t = w_{t-1}p(y_t|x_t) \quad (12)$$

Although this traditional particle filtering technique has limitations, in particular with regards to the description of the distributions tails, and more advanced resampling schemes have been proposed (Acuña & Orchard, 2017), this technique was still deemed valid for a purely preliminary analysis. Long-term prediction of the fault evolution can be obtained by iterating the "prediction" stage, and are used to estimate the probability of failure in a system given a hazard zone that is defined via a probability density function with lower and upper bounds for the domain of the random variable, denoted as H_{lb} and H_{up}, respectively. Given the probability of failure, the RUL distribution for any given prediction can be computed along with the risk function (Acuña & Orchard, 2018). The declination of the particle filter employed in this paper is based on a physics-based degradation model and a process model describing the dependency between the worn-out thickness x of the brake pads and the selected features.

$$\begin{cases} x_{N+1} = K_{wear}(E_{brake_N}) + x_N + \omega(N) \\ y_{N+1} = f(x_{N+1}, \nu(N)) \end{cases} \quad (13)$$

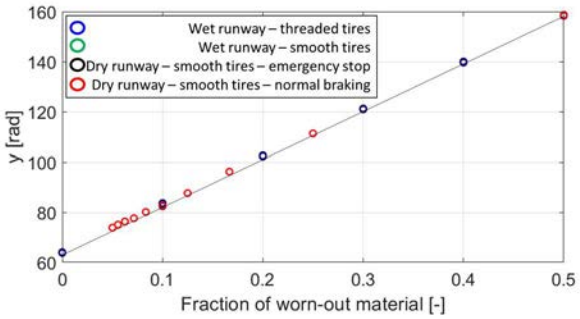


Figure 9. Dependency of the proposed feature on external factors and on degradation progression.

where K_{wear} is the wear constant, y is the feature and E_{brake_N} is the gross energy produced during the N^{th} landing. $\omega(N)$ and $\nu(N)$ are noises, estimated at each time step considering the probability distributions of the parameter and the accuracy of the process model through a certain number of previous steps.

The gross energy E_{brake_N} is estimated as follows, and provides an indication of an energy proportional to the terms of the modified Archard's Equation provided in Equation 7, following the expression:

$$E_{brake_N} = r_{ebrake} \int_{t_0}^{t_{end}} \sum_{i=1}^{4} F_i\,\omega_w dt \quad (14)$$

where F_i is the force exerted by each actuator and ω_w is the wheel angular frequency. The feature y is defined as the average angular position of the EMAs motor for which the measured force signal is significantly different from zero at the beginning of the braking procedure. Such data point is identified computing the rolling variance of the force signal and searching for the first point which rolling variance exceed 0.1. Such feature was successfully identified as the most promising in (De Martin, Jacazio, Parisi, et al., 2022) due to its high correlation with the degradation process. As shown in Figure 9, such feature is also not affected by variations in the operating conditions, nor by the braking procedure (emergency or normal). For each landing, the quantity E_{brake} is computed and memorized in a "landing repository", where it is stored along with related aircraft-level information, such as the aircraft weight at landing, for future usage.

During the long-term prognosis, the "landing repository" data-base is used to build an array of possible future landings through random sampling. If an indication or prevision of the area in which the aircraft is going to typically operate is available, a planned feature is to further refine the sampling procedure to account for the most probable weather conditions. The prognostic algorithm is tested against 40 simulated fault-to-failure processes, where the wear of the brake pads evolves dynamically as a function of the system behavior and operating conditions (temperature, dynamic load, fluid pressure), with increasing number of particles N_p

(from 50 to 5000) and evaluated according to the traditional metrics provided by (Saxena et al., 2008), namely the Prognostic Horizon, evaluated as the first real RUL value for which the prognosis falls within a $\pm$ 20% threshold of the real RUL, and the Relative Accuracy RA, defined as a function of the ground-truth value of the RUL (RUL_r) and its expected value RUL.

$$RA = 1 - \frac{|RUL_r - RUL|}{RUL_r} \quad (15)$$

An example of the prognostic output for the case of a is provided in Figure 10, where the system behavior is plotted against the number of simulated landing procedures N_L.

The behaviour of the particle filtering algorithm is investigated in two steps. At first considering the "filtering" performances, thus evaluating whether the system is able to correctly assess the severity of the on-going degradation, and secondly considering the long-term prognostic capabilities. Figure 11 depicts the behavior of the particle filter algorithm with respect to the simulated ground truth for the landing sequence already used for Figure 10, evidencing the particles distribution considering the estimated values assumed by the hidden state (the linear measure of the brake pad wear progression) and the selected feature. This information is given considering four equidistant prediction instances, with indication of the considered simulated landing.

It can be observed that the results of the particle-filtering routine are compatible with the simulated ground-truth in all of the shown cases, highlighting that the algorithm is able to coherently track the fault growth from the fault detection until imminent failure conditions. Figure 12 and Figure 13 describe the algorithm behavior against the simulated ground truth from a prognostic perspective. The estimated RUL distribution are coherent with the ground truth, and achieve convergence towards the simulated end-of-life.

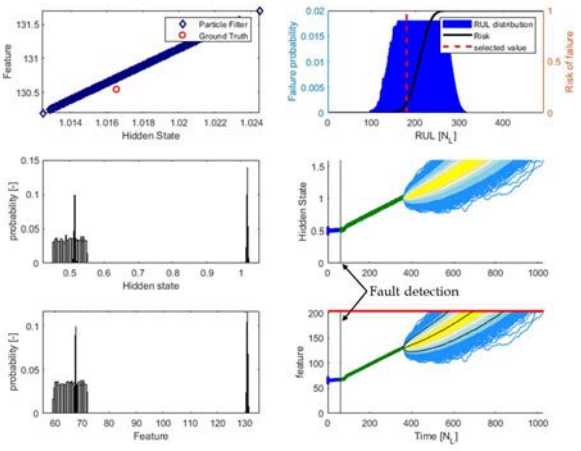


Figure 10. Prognostic performance against simulated data-set ($N_P = 5000$)

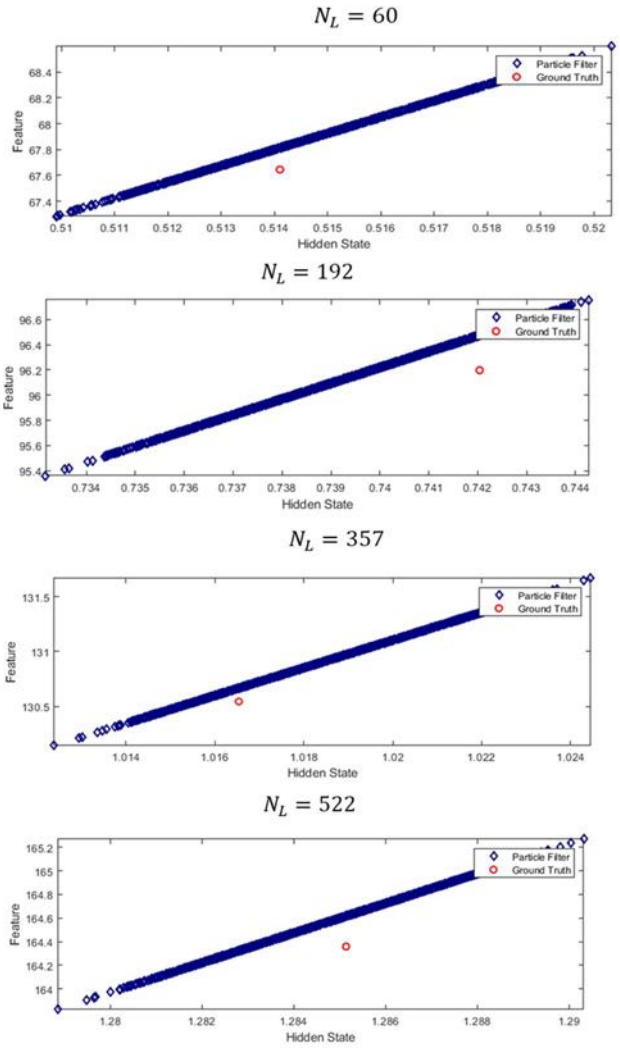


Figure 11. Comparison between PF and simulated ground-truth during the filtering stage ($N_P = 5000$)

In Figure 12, the RUL distribution at each considered prediction step is depicted along the selected value, corresponding to the RUL estimate with the highest probability of occurrence according to the algorithm. The "ground-truth" EOL, coming from simulation data set, is also provided. Results shows that the real EOL always falls within the prediction distribution, in the near proximity of values of risk of failure equal to 1. Although providing only anecdotal evidence – a more rigorous approach would be to compare the predicted RUL distribution against a real RUL distribution – this figure attests that the algorithm converges to the EOL in the analyzed case, providing then promising results.

This observation is confirmed by the $\alpha - \lambda$ diagram in Figure 13. The small deviation of the expected RUL with respect to the simulated ground-truth close to the EOL is expected to the prediction uncertainty increasing relatively to the RUL estimate.

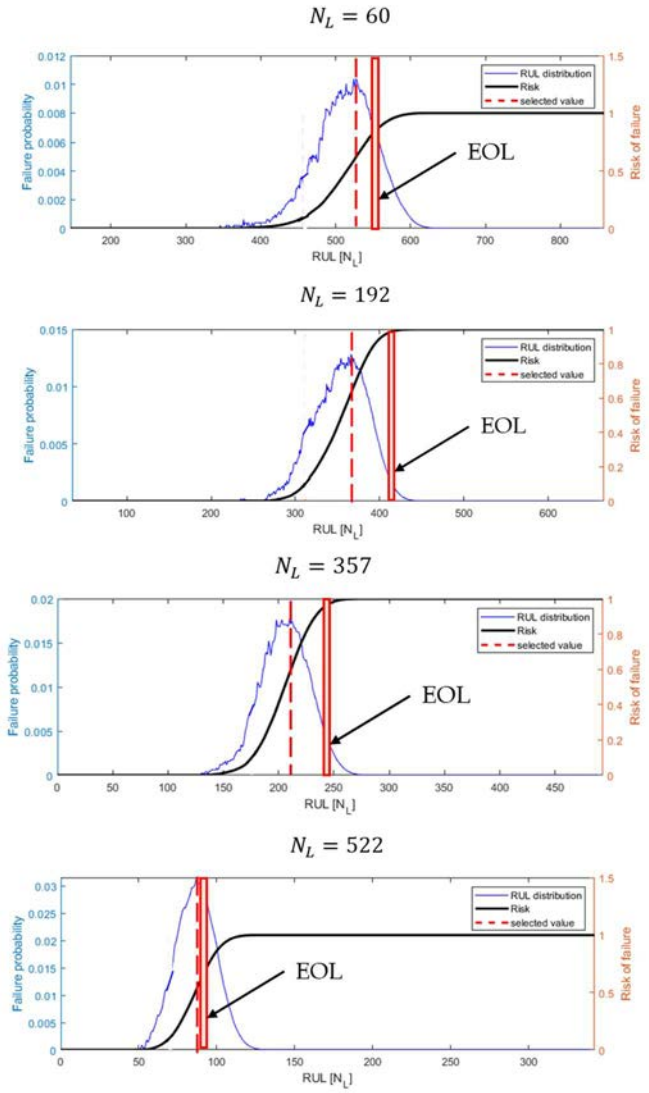


Figure 12. Comparison between estimated RULs at different prediction steps ($N_P = 5000$, $EOL = 600$ landings)

The prognostic performances of the algorithm are presented in terms of Relative Accuracy and Cumulative Relative Accuracy in Figure 14, where the results averaged over the 40 simulated landing sequences. It can be observed that the average Relative Accuracy remains well above the 80% threshold, while scoring high marks in CRA as well.

Finally, the algorithm is evaluated considering its elapsed time, to verify whether it is suitable for on-board or on-line deployment. Results were obtained on a DELL Precision 3660 workstation with *IntelCore i9-4.2 GHz* and *64GB* of DDR-4 RAM. Results are coherent with a possible real-time usage of the algorithm, although out of scope for the current project.

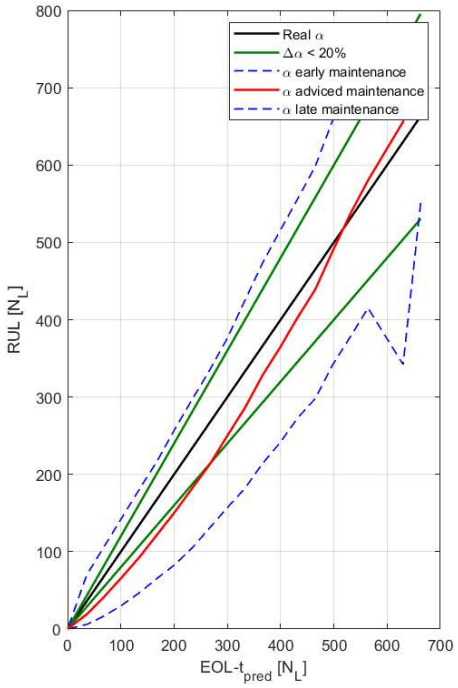


Figure 13. Dimensional $\alpha - \lambda$ diagram

6. PHM Implementation and Test Plan

Given the intermittent nature of the E-Brake operations, PHM algorithms are expected to be run offline, without the need to suffice restrictive computational constraints to achieve a real-time identification of the fault occurrence or progression. After each landing data are collected and analyzed. These same considerations are carried to the technological demonstrator, which is thought to monitor the E-Brake health status considering the results coming after each test.

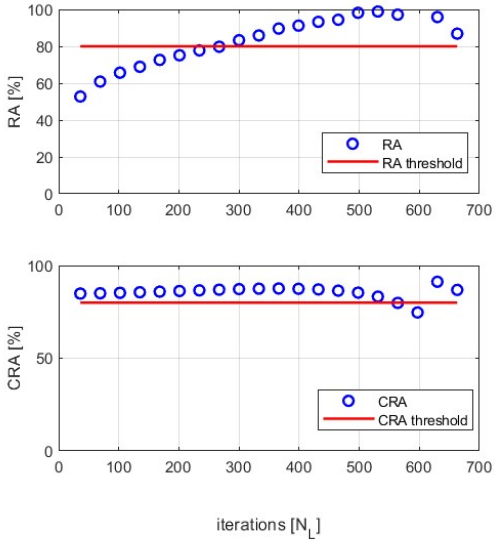


Figure 14. Prognostic performances (RA, CRA)

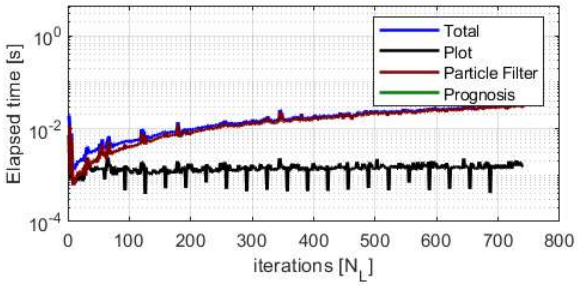


Figure 15. Algorithm elapsed time

The test plan is designed to verify and validate the most critical aspects of the presented work, namely the high-fidelity model employed to generate the synthetic database used to prepare the PHM activities and the wear model itself. Similarly, it is necessary to stress the PHM algorithm to check for false alarms and evaluate the performance of the prognostic output.

For these reasons the test plan includes:

- Registration of the qualification tests, including the simulated landing and the aborted take-off to evaluate the model behavior, considering different aircraft mass as detailed in Section 4.2.
- A number of consecutive landing simulations (at least 100), considering different aircraft weight and braking operation type (emergency, normal).
- The visual evaluation, after each landing, of the amount of worn thickness from the brake pads.

Since the test rig is not yet operational authors are not able to present the results of these activities, that will be hopefully the subject of further dissemination once the experimental steps are completed.

7. CONCLUSIONS

This paper presents the design of a novel technological demonstrator for PHM activities on a fully electrical landing gear and the preliminary design of a prognostic routine to forecast the wear in the pads of an electro-mechanical brake for a short-range aircraft. Since the test-rig is not yet operational, PHM activities have been tentatively proposed through a high-fidelity simulation model, iterating on the results of a previously published study. Such model, presented in the paper, is based on well-known equations, and translated into a state-of-the-art dynamic simulation engine. The PHM scheme is based on equations strictly correlated with the physics of the investigated degradation, and leverages the previous knowledge of the system usage to forecast the long-term estimate of the brake pads RUL. Early results are encouraging, but experimental support is needed to validate the findings of the simulation activities.

Further work will include the description and analysis of the effects of disks wear and the definition of health monitoring schemes to detect and prognose other prominent failure modes potentially affecting electrical brakes.

REFERENCES

Acuña, D. E., & Orchard, M. E. (2017). Particle-filtering-based failure prognosis via sigma-points: Application to Lithium-Ion battery State-of-Charge monitoring. *Mechanical Systems and Signal Processing*. https://doi.org/10.1016/j.ymssp.2016.08.029

Acuña, D. E., & Orchard, M. E. (2018). A theoretically rigorous approach to failure prognosis. *Proceedings of the 10th Annual Conference of the Prognostics and Health Management Society 2018 (PHM18), Philadelphia, PA, September 24-27*.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2009). A Tutorial on Particle Filters for Online Nonlinear/NonGaussian Bayesian Tracking. In *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*. IEEE. https://doi.org/10.1109/9780470544198.ch73

Bertolino, A. C., De Martin, A., Jacazio, G., & Sorli, M. (2023). Sizing and control system definition of an intelligent facility for qualification tests and prognostic research activities for electrical landing gear systems. *Materials Research Proceedings*, *26*, 219–224. https://doi.org/10.21741/9781644902431-36

Carbone, G., & Putignano, C. (2013). A novel methodology to predict sliding and rolling friction of viscoelastic materials: Theory and experiments. *Journal of the Mechanics and Physics of Solids*, *61*(8), 1822–1834. https://doi.org/10.1016/j.jmps.2013.03.005

De Martin, A., Jacazio, G., Parisi, V., & Sorli, M. (2022). Prognosis of Wear Progression in Electrical Brakes for Aeronautical Applications. *PHM Society European Conference*, *7*(1), 329–337. https://doi.org/10.36001/phme.2022.v7i1.3353

De Martin, A., Jacazio, G., Ruffinatto, A., & Sorli, M. (2022). A Novel Hydraulic Solution to Simulate Inertial Forces on a Landing Gear Qualification Test Rig. *Proceedings of the ASME/BATH Symposium on Fluid Power and Motion Control, FPMC2022*, 1–9.

De Martin, A., Jacazio, G., & Sorli, M. (2022). Simulation of Runway Irregularities in a Novel Test Rig for Fully Electrical Landing Gear Systems. *Aerospace*, *9*(2), 114. https://doi.org/10.3390/aerospace9020114

Giannella, V., Baglivo, G., Giordano, R., Sepe, R., & Citarella, R. (2022). Structural FEM Analyses of a Landing Gear Testing Machine. *Metals*, *12*(6). https://doi.org/10.3390/met12060937

Grosso, L. A., De Martin, A., Jacazio, G., & Sorli, M. (2020). Development of data-driven PHM solutions for robot hemming in automotive production lines. *International Journal of Prognostics and Health Management*, *11*, 1–13.

M. Burckhardt. (1993). Radschlupf-Regelsysteme. *Fahrwerktechnik: Würzburg: Vogel Verlag.*

Mohan, N., Undeland T.M., & Robbins, W. P. (2005). *Power Electronics* (3rd ed.). John Wiley and Sons, Inc.

Oikonomou, A., Eleftheroglou, N., Freeman, F., Loutas, T., & Zarouchas, D. (2022). Remaining Useful Life Prognosis of Aircraft Brakes. *International Journal of Prognostics and Health Management*, *13*(1), 1–11.

Olesiak, Z., Pyryev, Y., & Yevtushenko, A. (1997). Determination of temperature and wear during braking. *Wear*, *210*(1–2), 120–126. https://doi.org/10.1016/S0043-1648(97)00086-0

Orchard, M. E., & Vachtsevanos, G. J. (2009). A particle-filtering approach for on-line fault diagnosis and failure prognosis. *Transactions of the Institute of Measurement and Control*. https://doi.org/10.1177/0142331208092026

Ramesh, G., Garza, P., & Perinpanayagam, S. (2021). Digital simulation and identification of faults with neural network reasoners in brushed actuators employed in an e-brake system. *Applied Sciences (Switzerland)*, *11*(19). https://doi.org/10.3390/app11199171

Roemer, M. J., Byington, C. S., Kacprzynski, G. J., Vachtsevanos, G., & Goebel, K. (2011). Prognostics. In *System Health Management: With Aerospace Applications*. https://doi.org/10.1002/9781119994053.ch17

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. *2008 International Conference on Prognostics and Health Management*, 1–17. https://doi.org/10.1109/PHM.2008.4711436

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). Intelligent Fault Diagnosis and Prognosis for Engineering Systems. In *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470117842

Yevtushenko, A., Kuciej, M., & Topczewska, K. (2017). Analytical model for investigation of the effect of friction power on temperature in the disk brake. *Advances in Mechanical Engineering*, *9*(12), 1–12. https://doi.org/10.1177/1687814017744095

Development of Anomaly Detection Technology Applicable to Various Equipment Groups in Smart Factory

Kiwon Park[1], Myoung Gyo Lee[2], Sung Yong Cho[3], Yoon Jang[4], Young Tae Choi[5]

[1, 2, 5] Hyundai Motor Company, 37, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do, Republic of Korea
king521dom@hyundai.com, hesse667@hyundai.com, 9562768@hyundai.com
[3, 4] Kia Motors, 37, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do, Republic of Korea
sungyongcho@kia.com, yunifree@kia.com

ABSTRACT

This study delves into the creation of anomaly detection technology applicable to a range of equipment groups within smart factories. This advanced technology uses high-performance MEMS vibration sensors, edge CMS devices, and PHM platforms to tackle issues such as data imbalance, learning model limitations, complex equipment operating patterns, and real-time processing. It also addresses central server concentration, data cycling problem, various equipment classification, and algorithm operation problems that can arise when implementing systems in the field. Using AI-based vibration detection algorithms, data can be collected at high sampling rates and analyzed in real-time through edge computing, minimizing latency and mitigating server capacity issues compared to cloud-based analytics. The system continually monitors and learns standard performance data from equipment to provide practical solutions that minimize equipment failures and downtimes. The results of this study are impressive, as it has successfully developed anomaly detection framework and PHM systems that are expected to enhance the efficiency and sustainability of smart factories. Furthermore, the study aims to showcase and improve the effectiveness of predictive maintenance in both domestic and international automotive factory production lines. This revolutionary technology will be a key component in smart and software-defined factories and help companies achieve intelligent automation.

1. INTRODUCTION

The rise of smart factories has recently led to an increase in production line automation equipment. As a result, maintenance activities have become crucial, and the need for predictive maintenance technology that can foresee equipment failures has emerged. Many companies are exploring ways to perform predictive maintenance, from installing additional sensors to analyzing controller data. Currently, predictive maintenance technology is limited to equipment that moves at a constant speed, like large turbines and fan motors.

We have developed a PHM (Prognostics and Health Management) system and an AI-based vibration detection algorithm capable of predicting anomalies in constant and variable-speed equipment to meet this need. Our technology stands out as it can collect vibration data at a high sampling rate, perform AI learning, and make decisions at the edge.

The PHM system consists of two primary components: the CMS (Condition Monitoring System) module and the PHM platform. The CMS module is a device equipped with edge computing functions, data collection capabilities, and decision-making algorithms. The PHM platform, on the other hand, monitors mining data from the CMS module, manages its operations, and deploys registered algorithms as a service for each CMS. Additionally, the platform is responsible for deploying the optimal algorithm as a module. The algorithm used in the PHM platform is developed through deep learning AI modeling and is registered and deployed as a CMS module [1].

Four-stage research was undertaken to create an AI-powered anomaly detection algorithm that relies on vibration data. Initially, a PoC (Proof of Concept) scenario was devised that focused on identifying the target equipment (robot reducer, automation equipment drive motor, etc.), the operating type (constant/variable speed, part/finished product), and the target defect type (robot reducer defect, motor bearing damage, etc.). Next, data was collected based on specific criteria for data type (vibration, current, speed, etc.) and collection method (CMS module, PLC, Cloud, etc.). In the third stage, signal feature extraction methods were defined through feature-based analysis, which uses domain knowledge to determine data analysis. An anomaly detection method was also developed to check abnormal scores by learning the normal group to suit the data imbalance situation where it is challenging to secure abnormal data compared to normal. The AI model used an Auto-Encoder structure and an unsupervised learning method, and an optimal model was developed through hyper-parameter adjustment to define the anomaly score [2]. The algorithm was verified through PoC activities by matching the score of the normal/abnormal state

of equipment with the actual motor defect phenomenon.

The PHM system and AI anomaly detection algorithm operate within the production line to learn and monitor the equipment's standard performance data. An alarm prompts maintenance activities when a score falls outside the normal range [3]. This approach minimizes equipment failure and ultimately aims to reduce non-operation rates. By leveraging data analysis to inform condition-based maintenance activities, our system establishes highly efficient maintenance and production plans that surpass traditional time and usage-based approaches. As a result, we can lower costs associated with non-operation rates within the production line [4].

2. BACKGROUND

Current predictive maintenance systems have typically utilized centralized models to analyze and predict facility-level data gathered from a central server. This approach has allowed rule-based algorithms to successfully extract key characteristics and implement predictive maintenance, even in low data sample rate environments where the facility operates at a constant speed.

However, operation patterns have become more complex with the rise of smart factories and diverse automation equipment. As the number of transmission equipment containing acceleration and deceleration patterns increases and the data types become more varied, data quality and accuracy have become increasingly important. A high data sample rate is required to analyze these transmission facilities effectively, and the utilization of AI algorithms has become crucial.

As a solution, we have developed a cutting-edge PHM system that seamlessly integrates an edge device and platform. With the power of edge computing technology, this system can efficiently process data near the facility, thereby reducing the volume of data transmitted to the central server. This not only lessens network load but also alleviates server burden. Furthermore, the system's real-time processing capabilities have been enhanced, resulting in a faster response time for our predictive maintenance system.

A framework for detecting anomalies that utilize advanced AI algorithms has been crafted to manage high data sample rates and identify significant features for precise anomaly detection. This framework has been customized for different kinds of facilities and effectively fulfills the requirements of smart factories.

Adopting smart manufacturing has necessitated a departure from conventional, centralized PHM systems towards decentralized, intelligent, and adaptive solutions. Incorporating edge computing and robust AI analytics is a forward-looking measure that promotes operational efficiency and dependability in contemporary automated facilities. Such innovations react to the evolving industrial landscape and a deliberate strategy to harness sophisticated technologies for more accurate fault prediction and prevention.

3. CHALLENGES

3.1. Challenges with PHM system developments

3.1.1. Data imbalance

Smart factories primarily collect steady-state data, which poses a challenge in detecting anomalies. The lack of abnormal data makes developing effective anomaly detection models difficult, as supervised learning models require sufficient labeled abnormal data. However, intentionally creating abnormal states in real environments is not feasible [5]. As a solution, an experimental test bench can be created to simulate normal and abnormal conditions to ensure a continuous supply of abnormal data. This data can be used to perform PoC verification. By conducting PoC, we can collect normal/abnormal data based on test conditions and create labeled data for each facility. This enables the use of highly accurate supervised learning models.

Unsupervised or semi-supervised learning methods are commonly used to solve the data imbalance in in-line. Unsupervised learning uncovers hidden patterns without labels, while semi-supervised learning enhances model performance by utilizing limited labeled data. This approach leverages smart factory inline data, primarily normal data, to establish normal distribution benchmarks for monitoring status. We can monitor anomaly score set up the lines divided warning and fault. By configuring and implementing the system on the production line, we can effectively address issues related to data imbalances.

3.1.2. Challenges with learning models

While unsupervised or semi-supervised learning methods can effectively identify anomalies, they do have a drawback because it can be challenging to pinpoint the exact cause of the anomaly. For instance, if a model detects an abnormality, it does not necessarily reveal whether the sensor responsible for the anomaly is faulty. To address this issue, it is necessary to conduct a thorough re-analysis of the facility's data after an anomaly is detected. Data from sensors must be separated and examined individually for each moving part or component location in the equipment, and specific patterns or characteristics contributing to the anomalies must be identified. To accurately determine the characteristic frequency based on the rotational speed of each component and identify any unusual fluctuations, we utilize domain knowledge to collect statistical data on gear frequency bands prior to the learning process. This stored information can be easily accessed through our platform for thorough analysis. This approach can be incredibly helpful in taking practical measures to resolve the issue.

3.1.3. Equipment challenges with complex patterns of robots

Sophisticated machinery, such as robots, can be challenging to monitor for irregularities due to their frequent acceleration, deceleration, and complex patterns. Additionally, the lengthy cycle times and diverse movements of robots make it difficult to identify patterns using traditional methods. However, advanced algorithms, including cycling techniques and signal processing methods such as time-frequency transformation (STFT) [6], can be applied to more accurately detect abnormal changes. These techniques make it possible to detect abnormalities with a higher degree of accuracy, even in facilities with complex patterns.

3.2. Challenges when applying PHM in the field

3.2.1. Challenges with the central server concentration method

In a cutting-edge factory setting, copious amounts of data are produced from diverse facilities. Specifically, intricate automation facilities generate significant quantities of data, including high sample rate vibration data. However, the conventional approach of transmitting this data to a central server for processing results in heightened network load and latency [7]. A practical solution to this challenge is to create an 'edge + platform system' leveraging edge computing technology to analyze data in real-time near the facility, extract critical information, and transmit it to the central server. This approach can expedite data processing while minimizing communication costs and central server storage management cost.

3.2.2. Challenges with cycling

In order to analyze data, it is necessary to cycle the data for a certain period, and PLC data is often used for this purpose. PLC typically utilizes line start/end process signals to timestamp data accurately. While low sample rate data is easily timestamped, high sample rate data like vibration presents a challenge. While the existing method to set timestamps was straightforward given the low data sample rate, more intricate equipment necessitates using high-sample rate vibration data to prevent information loss through down-sampling. However, this presents a challenge when attempting to set timestamps with PLC due to its limitation of about ten samples per second to avoid taxing the controller. As vibration sample rates can reach up to 16 k Samples/second, the resulting difference of approximately 1600 times is sufficient for information loss. A cycling or robust delay learning method is needed to address this issue.

3.2.3. Challenges with algorithm operation

Given the dynamic nature of smart factory environments, ensuring that AI algorithms remain up-to-date is crucial. To achieve this, an MLOps must be implemented, enabling periodic retraining and redeployment of algorithms [8]. Moreover, a collaborative approach between operating departments and maintenance organizations must be established to adapt swiftly and effectively, with a mechanism in place for rapid feedback and adaptation.

The upkeep and enhancement of AI algorithms demand consistent attention and a well-structured approach. To achieve this, the operations, conservation, and AI development departments must collaborate closely. Through monitoring data, identifying areas that require retraining, and leveraging real-time operational feedback, they can optimize the predictive conservation system's performance, leading to heightened efficiency.

3.2.4. Challenges with directing maintenance workers

In the early stage of system implementation, maintenance workers may face challenges in responding promptly to fault alarms. To enhance their response effectiveness, it is imperative to set up a system that showcases the blueprint of each facility on the platform and highlights the precise location of the alarm. The platform clearly presents the factory layout, indicating the location of all facilities. Every moving part of the facility has sensors and edge devices placed in precise locations, making it easy for maintenance workers to identify maintenance work locations through alarms displayed on the screen. Achieving this level of efficiency should be effortless. This approach will significantly boost the speed and precision of maintenance work.

4. METHODS

4.1. PHM System

4.1.1. Vibration sensor

Smart factory transmission equipment requires more precise and accurate data analysis. For this purpose, we used a vibration sensor with a sensitivity of 160 mV/g and a sampling rate of more than 8 k Samples/second to obtain high-quality data. These high-performance wired vibration sensors have the disadvantage of incurring additional installation costs. Still, applying a cost-effective, inexpensive vibration sensor of the MEMS type has compensated for this disadvantage. This makes it possible to collect high-quality vibration data economically.

Table 1. Vibration sensor

Vibration Sensor	Type	MEMS
	Axis	Z (mono)
	Sensitivity	160mV/g

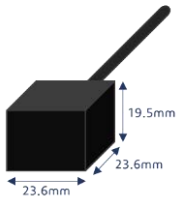


Figure 1. Vibration sensor

4.1.2. Edge CMS (Continuous Monitoring System)

As the amount of data increases exponentially, concentrating data on a central server for analysis becomes difficult due to storage capacity management and data latency issues. In particular, high sampling rate data processing is essential in facilities with complex patterns, which makes centralized analysis more difficult. To respond to this, we developed Edge CMS with edge computing capabilities, processing high-sample rate vibration data in real-time at the edge, extracting features, and calculating AI scores. The system supports 24-bit resolution and a sampling rate of 16kS/s, allowing processing without data loss. AI algorithms are mounted on these Edge CMSs and can make decisions immediately near the facility.

Table 2. Edge CMS module

CMS Module	Max. Sampling Rate	16 k sample/sec
	Channel	8 channel
	Bit Resolution	24 bit

Figure 2. Edge CMS module

4.1.3. PHM Platform

The PHM platform is located on the central server and manages each Edge CMS device applied to each facility. The AI algorithm is registered in the platform as learned and then distributed to the CMS, which requires updates when necessary to optimize and manage abnormality detection. For example, if the line situation changes and the operating pattern teaching is modified, two weeks' worth of raw data is relearned, and the learned model is redistributed to the CMS located in the relevant process facility for operational management. In other words, the MLOps cycle that allows re-learning/re-distribution was implemented. Key features and AI scores calculated from the CMS located at each facility are transmitted to the platform and displayed to check trends by date. If the appropriate standard value is exceeded, a warning and fault alarm is given to notify the operator, and it displays which equipment and location on the layout shows signs of abnormality, helping to instruct maintenance workers on maintenance work. The platform layout was modeled after each factory line, and the web screen was designed so that if an error occurs, the area is marked in red to be visually checked immediately.

4.2. Anomaly detection framework

4.2.1. Cycling Techniques

'Cycling' is 'Extracting one cycle in operating data patterns of equipment'. Our data must be 'cut off' in the equipment operation cycle. Usually, cutting is done with PLC signals, but down-sampling is necessary to match the start / end signal timestamps to the data. However, the simple down-sampling method may be ineffective when collecting vibration ineffective when collecting vibration data at 8kS/s

Figure 3. PHM platform monitoring screen composition

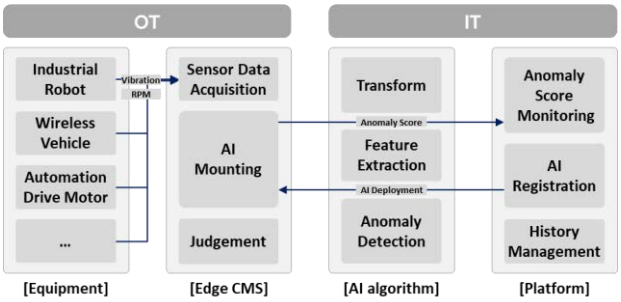


Figure 4. A configuration block diagram of the PHM platform and Edge CMS module connection with AI algorithm in the OT/IT range

to analyze transmission equipment that repeats acceleration/deceleration. To overcome this, we utilized an auto-cycling technique to divide the acceleration, constant speed, and deceleration sections. We obtained the specific frequency by determining the rotational frequency based on the equipment motor's RPM. We then counted the peaks of the acceleration/deceleration in both time and frequency and set a vibration magnitude threshold to divide the acceleration/constant speed/deceleration sections. This technique can be applied to various equipment motors, including lifts, conveyors, and stackers/destackers for transporting logistics boxes or vehicles. Regardless of the distance traveled, the acceleration/constant speed/deceleration types can be learned and utilized separately.

Advanced gear-shifting equipment, such as robots, faces a challenge when splitting acceleration and deceleration using auto-cycling techniques. As a result, the entire one cycle must be used for learning. The process signals receive start and end bits, which are then used to set the cycling point. To ensure robust learning despite delays caused by different sample rates, features are imaged at a later point. The frequency distortion caused by converting the entire cycle is resolved through STFT conversion, allowing for the utilization of all time-frequency information.

4.2.2. Preprocessing and Conversion

Data value can vary depending on the unique conditions of each facility. To refine normal data, checking its distribution, applying DC offset, and filtering where necessary is essential. Since most equipment comprises motors and gears, confirming rotation frequency in the frequency spectrum based on speed is possible. Features are extracted to ensure accurate expression of rotational frequency and harmonic components based on gear mesh theory [9], and window size is set to perform FFT spectrum conversion up to the 4kHz band [10], [11]. After conversion, RMS statistics are calculated for each harmonic frequency band and basic statistics like Min, Max, Average, Kurtosis, and Skewness [12]. This data is stored on a server to enable detailed analysis and monitoring. The magnitude of the FFT spectrum converted to a 1D shape is fed into the AI algorithm for further study. The spectrum is transformed with a window size of 3 seconds, and data is extracted through window sliding with a duration of 0.25 seconds [13]. In the

slow-speed equipment, the spectrum is reduced to 1.5 kHz, and the conversion values for each channel are concatenated and studied in the form of a wave set. This process has led to the development of an optimized anomaly detection framework that applies different conversion techniques to suit the specific characteristics of each facility.

Robots utilize STFT transformation to extract features, which involves cycling in the manner described above. The output of STFT is a 2D shape from a colormap image, which serves as input to the AI algorithm. When features are extracted using 1D Conv, some degree of conversion freedom allows for flattening and use within the algorithm. The STFT value is also stored separately and used for detailed feature analysis. The robot stores features in separate channels for each axis to allow for more accurate analysis. This approach enables the identification of any anomaly score increase in a specific channel, which can then be used to issue maintenance instructions for the affected axis. The FFT spectrum is re-extracted for the robot's statistics, using a window size of 3 seconds within one cycle. The extracted features are stored similarly to the driving motor of general equipment and are not separately learned by the AI model. They are stored on the server for monitoring during detailed analysis.

4.2.3. AI algorithm

The most effective method for confirming data classification is Dimensional Reduction Visualization. This involves reducing the extracted features' dimensions and representing them on a 2D graph's x and y axes. Doing so can ascertain how the feature distribution is formed by date and whether it is clustered. LDA (Linear Discriminant Analysis) is utilized for dimensionality reduction [14]. The average feature value for each date is represented as a single point, and each month is color-coded to show how the features change visually from one month to the next.

Supervised learning struggles to classify typical smart factory data due to the difficulty of obtaining abnormal data. However, clear labeling can ensure the accuracy of this method. To address this issue, we developed an algorithm to collect abnormal data and classify the collected abnormal data so that it can be distinguished from normal data in various scenarios, such as motor misalignment, bearing failure, bearing cage damage, robot reducer failure, and lubricant shortage. We utilized deep learning, specifically a convolution method, to capture features easily using spatial information of image data. The 1D CNN layer consisted of 3 layers, utilizing the relu activation function [15], a model was created to classify into normal/abnormal through the dense layer. After verification, over 97% of the classifications were confirmed. The classification model was saved in the system, and data on diverse types of defects were collected and attributed to the system for future classification purposes.

In cases where there is insufficiently abnormal data, Anomaly Detection can be achieved by establishing a baseline of what constitutes "normal" data and monitoring any deviations from that baseline through a scoring system. An AI algorithm utilizing an auto-encoder structure must be introduced to employ this method [16]. Before training the model, extracted features are inputted as the model's input values. When dealing with robots, input values take the form of images, for which a convolution layer is created to facilitate image analysis. This layer consists of three layers with a relu activation function, and instead of pooling, it utilizes the stride technique to employ all pixel information [17]. The decoder comprises three Conv2DTranspose layers reconstructing the extracted features. Learning uses the Adam optimizer, mae loss, and appropriate batch_size and learning_rate parameters. The trained model predicts new data with the same feature shape, calculates the loss difference from the normal group learning value, and generates an anomaly score. Anomaly detection is achieved by monitoring this score over time and checking the platform display for gradual increases.

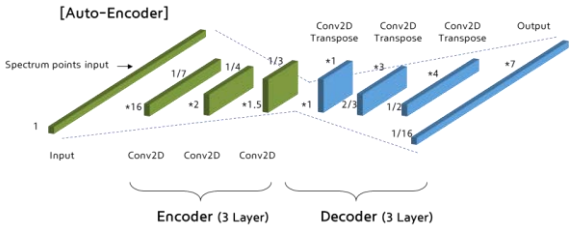


Figure 5. Configuration of vibration-based AI model (Auto-Encoder) for Anomaly detection

5. Verification

5.1 Verification of PoC

5.1.1 Construction of Test-bench and data collection environment through PoC

We set up an external test bench and conducted a PoC test to gather information on the target equipment. In the case of industrial robots, we installed vibration sensors on each axis of the reducer part for manufacturers such as Hyundai, ABB, Yaskawa, and Kawasaki. We repeatedly drove the machine by teaching it a complex, 6-axis movement that could withstand heavy use. To collect vibration data from the reducer part, we used a motor with the same capacity as the logistics line conveyor and lift equipment and connected a load system to apply a constant speed drive. We monitored and verified the target type by selecting equipment with a high non-operation rate in-line, such as when replacing a reducer due to mechanical defects.

5.1.2 Development of Motor PHM diagnosis algorithm and Verification in Test-bench

A dynamometer was installed on the motor and reducer (manufactured by SEW) to confirm the PoC, which drives the automation equipment utilized in actual mass production. A load was applied, and critical parts were equipped with a vibration sensor to collect data. The system was run at a constant speed of 1800rpm with load currents ranging from 2.5 to 0.1A.

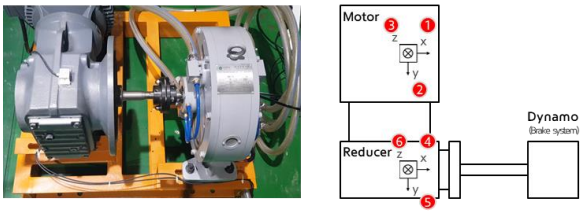


Figure 6. Real and dynamometer configuration diagram of Motor and Reducer

In our analysis, we compared normal and abnormal data. Specifically, we examined changes in vibration magnitude on the time axis and alterations in specific frequency values, band ranges, and harmonics of the rotating body on the frequency axis. Using this information, we conducted a date-wise assessment to determine if there was a gradual change.

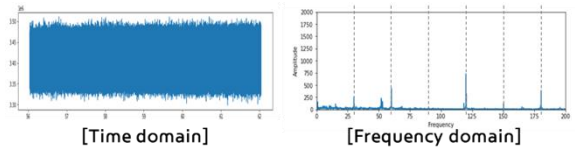


Figure 7. Normal data pattern by time/frequency domain

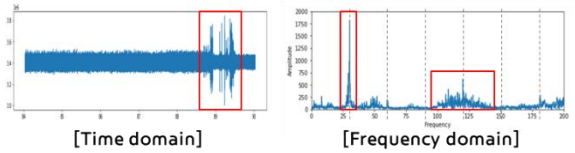


Figure 8. Abnormal data pattern by time/frequency domain

Our process involves extracting feature vectors from acquired data using FFT spectral transformation. These vectors are then input into an autoencoder model, which returns an output vector. The model first learns the vector of the normal group, compares it with the vector of new data, and returns the error value as the final score using MAE. To test this, we selected ten days of motor operation data with the same conditions and the occurrence time of an abnormality. We trained the model using the first three days and predicted the next seven days. Finally, we analyzed the predictions to identify any changes in the data.

Figure 9 displays vibration RMS values over time, indicating that the RMS only increased at the time of failure, making it challenging to predict through rule-based measurement value monitoring. However, in Figure 10, the Anomaly score gradually increases over time. Figure 11 displays the average value per date, revealing an increasing score from January 28th. As the anomaly score rises, data that differs from the standard norm is being collected, making it feasible to operate a PHM system that anticipates failure and notifies the time of failure via an anomaly score baseline of 0.2 to 0.3.

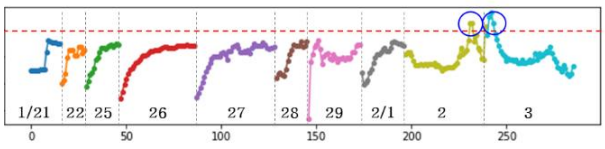


Figure 9. Comparison of vibration RMS values by date

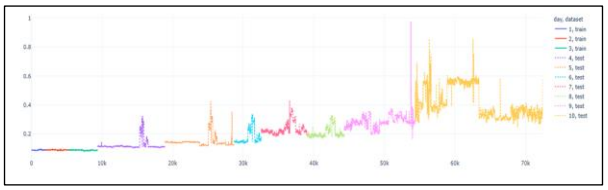


Figure 10. Comparison of Anomaly scores by date

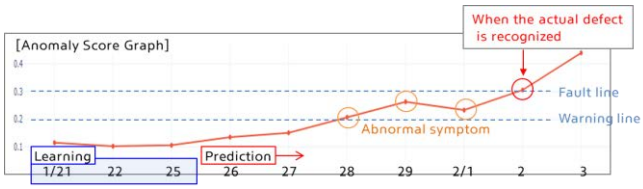


Figure 11. Comparison of Anomaly scores average

After the actual abnormal data was acquired, the operation stopped due to motor failure after continued operation for two months, and a motor disassembly analysis was performed to confirm the phenomenon.

The cause is damage to the reducer and internal bearing due to dynamometer misalignment. Damage to the bearing cage and excessive tooth surface wear can be seen in Figure 12. As a result, a prediction model for motor failure was developed, and it was confirmed that the algorithm could be applied and operated by matching the failure phenomenon.

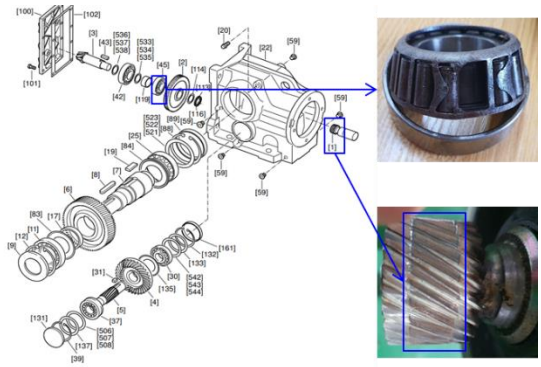


Figure 12. Result of disassembling bearing of faulty reducer

5.2 Verification of Production factory in-line

5.2.1 Inline data analysis process

An aging robot was selected from the in-line welding robots used for car body production to verify the development algorithm. A vibration sensor was installed, and data was collected. The robot was significantly aged after operating for 11 years without a reducer replacement. Upon analysis of the iron concentration in the reducer grease, it was found to be approaching the replacement criterion. The data was monitored for an additional four months, and the reducer was replaced with a new product. The change in Anomaly score was checked to verify the replacement. This verification process aims to determine whether the aging pattern is distinguishable from normal, even if it is not a failure, and whether the score can increase gradually and eventually lead to a failure.

Figure 13. Appearance of aging robot reducer in-line

The anomaly detection process utilized the Auto-Encoder model through AI technology. Upon replacement, the data was swiftly learned and compared to the data not learned before replacement using the AE model for score evaluation by date. Below is the comprehensive data analysis procedure that is verified based on what is described in Chapter 4.

Using raw vibration data, we extracted one-cycle data by analyzing the similarity or the on/off signals of a PLC. The data collected includes six channels and spans 51 seconds, with sensors installed on each axis of a 6-axis robot. However, the window size was too large to analyze one-cycle data in the frequency spectrum. To address this, we used partial data corresponding 1~3 seconds to transform to FFT spectrum or performed an STFT transformation using total cycled data to extract features in both time and frequency bands. This is a

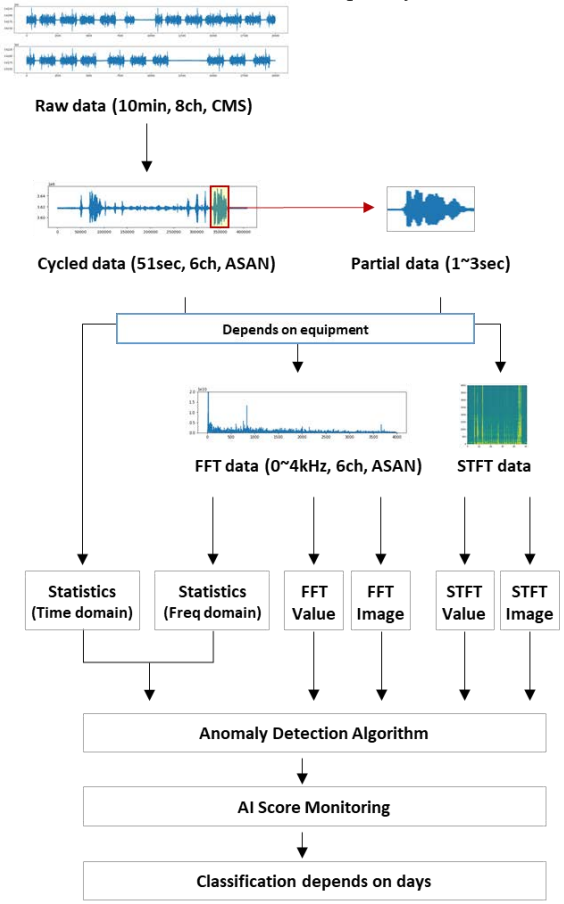


Figure 14. In-line data analysis process

key method for extracting features based on the equipment and operation pattern. The converted STFT was then shaped into an image with the following input dimensions.

Abnormal dataset: (580, 385, 387, 3)
Normal dataset: (185, 385, 387, 3)

To ensure accurate and reliable results, we split 60% of the data set into a training dataset of 459 samples and a test dataset of 306 samples. This allowed us to effectively organize and analyze the data before proceeding with the learning process.

5.2.2 Visualization of data distribution

To visually represent the distribution of data, we employed dimensionality reduction using the LDA (Linear Discriminant Analysis) technique. This involved breaking it down into two-dimensional components and displaying it on a 2D graph. M5 to M10 in the graph represent months. After the reducer was replaced, October was expressed in brown, and the months from May to September before the replacement were expressed in a different color. Upon observation, we concluded that the data's distribution clusters were formed differently before and after the reducer replacement.

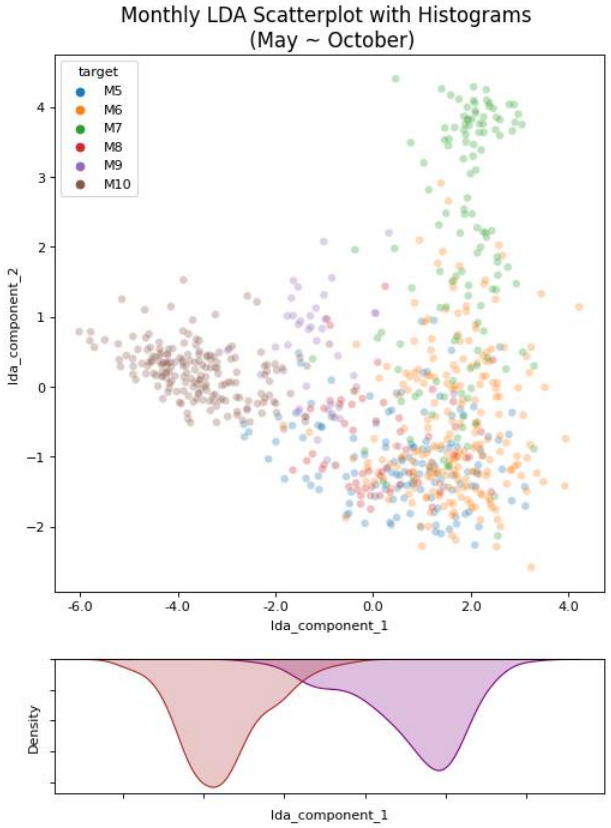


Figure 15. Monthly scatter plots and histograms were separated using Linear Discriminant Analysis – Abnormal (M5~9_purple) and Normal (M10_brown).

5.2.3 Results of Anomaly Detection Analysis

We use STFT colormap image as the characteristic feature. Utilizing colormap images demonstrated superior feature extraction through a convolution layer [18]. This led to a highly effective anomaly detection model, which relied on an Auto-Encoder structure as its foundation. Specifically, we constructed an encoder consisting of three convolutio

and a decoder featuring a convolution transpose layer.

50% of the normal dataset was used for learning, and the remaining abnormal and normal datasets were used as a test set to check the anomaly score.

A discernible visual difference was observed in the scatter plot after comparing the reconstruction error before and after replacement with a new product. The distribution graph distinguishes the abnormal state in orange color and normal state in blue color before and after the replacement time point. (Figure 17). Subsequently, upon setting the error threshold, the distinction in distribution between the normal state, which is represented in dark blue and the abnormal state, which is represented in orange, was confirmed through the histogram. These findings suggest that the replacement product had a significant impact on the reconstruction error and, thus, could enhance the overall performance of the system. (Figure 18)

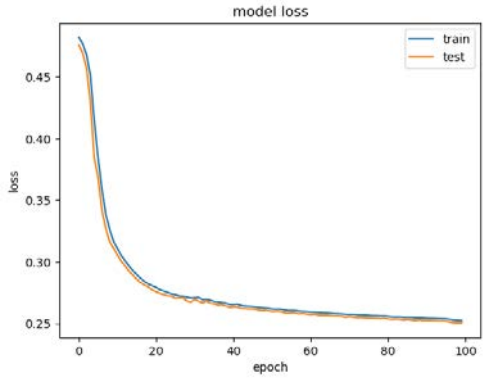


Figure 16. Auto-Encoder Model learning loss graph

Learning was performed by repeating epochs to minimize loss.

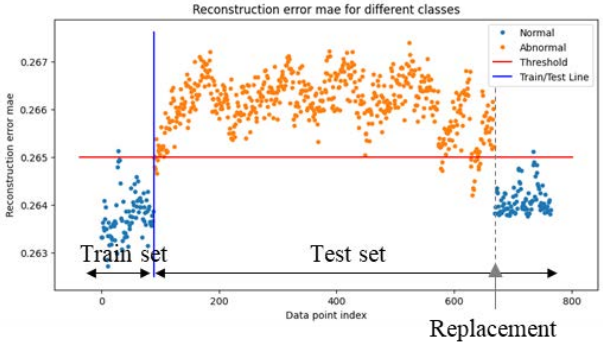


Figure 17. Reconstruction error scatter plot by Train/Test set

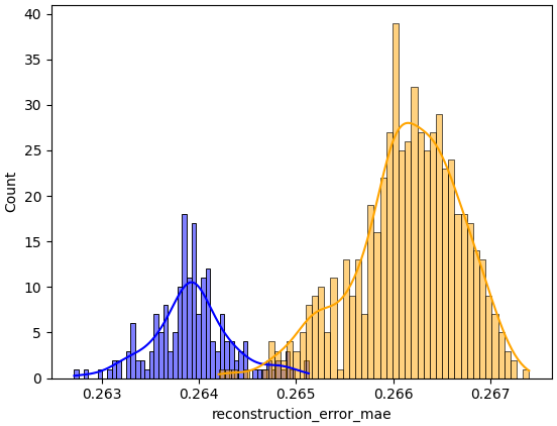


Figure 18. Reconstruction error distribution histogram

It was determined through predictive analysis of the test dataset that significant differences in reconstruction error exist between the normal and abnormal datasets. The formation of distinct clusters in the reconstruction error distribution further confirmed these differences. By setting the appropriate threshold value and checking the classification accuracy, it was ascertained that the classification was highly accurate, exceeding 97%. Additionally, by monitoring the anomaly score through unsupervised learning, variations in data patterns in aging equipment were identified as indicative of potential breakdowns.

Table 3. Test set classification results depending on metrics

Accuracy	Precision	Recall	F1	ROC_AUC
97.04%	99.82%	96.72%	98.25%	97.84%

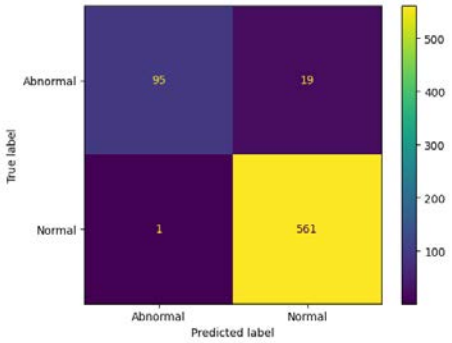


Figure 19. Ab/Normal state confusion matrix of the Test set

6. Smart Factory Application

Our new smart factory's assembly and logistics lines currently utilize the PHM system. It's applied to both constant-speed equipment like fan motors and variable-speed equipment like robots, lifters, and wireless mobile vehicles. Through our research and development, we've been able to internalize our technology and significantly reduce construction investment costs compared to external products. Moving forward, we plan to constantly monitor data and enhance our algorithm by identifying and addressing specific facility defects. Our ultimate goal is to expand horizontally, proving and verifying the effectiveness of predictive maintenance on domestic and overseas automobile factory

production line facilities.

6.1. Classification of Smart Factory Equipment Group

6.1.1 Constant speed equipment

When it comes to equipment that operates at a consistent speed, the relevant data values are monitored and analyzed, or specific window sizes are set to monitor the score obtained from learning features within the frequency spectrum. This applies to equipment like painting fan motors, supply/exhaust fans, and air blow pumps.

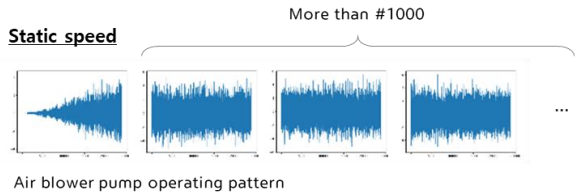


Figure 20. Operation pattern of constant speed equipment

6.1.2 Monotonic acceleration/deceleration pattern equipment

Regarding in-line equipment, movement automation equipment may follow a repetitive pattern of acceleration, deceleration, and constant speed. This is evident in stacker/de-stacker equipment that moves BIW between floors in an automobile manufacturing line and conveyor belt drives for movement between processes. Despite having variable-speed capabilities, the acceleration/constant speed/deceleration pattern remains constant, allowing for the extraction and utilization of features across the frequency spectrum by dividing the pattern accordingly.

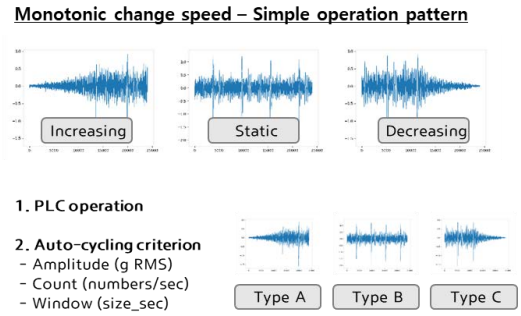


Figure 21. Operation pattern of acceleration/deceleration equipment

6.1.3 Monotonic acceleration/deceleration and various pattern equipment

In logistics box warehouses, repetitive acceleration and deceleration patterns are common, but the locations of the logistics boxes can vary greatly. This is particularly true with equipment such as the MSC (Multi Stacker Crane) and the SC (Stacker Crane). Given the vast number of possible box positions, it is not practical to learn the patterns of each position individually. Instead, the deceleration, constant speed, and acceleration patterns are divided, and learning is conducted by selecting features through the frequency spectrum.

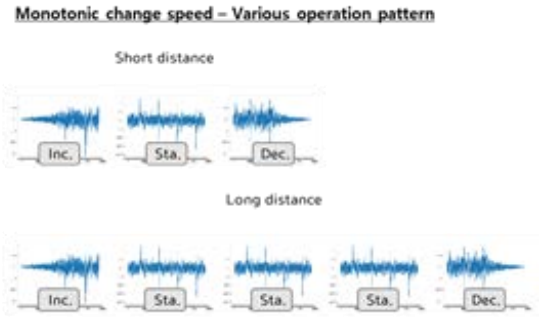


Figure 22. Operation pattern of wireless mobile device

6.1.4 Complex acceleration/deceleration pattern equipment

With the rise of smart factories, 6-axis industrial robots have become the primary choice for automation. However, due to the complexity of their multi-jointed structure, obtaining features for each movement can be challenging. To overcome this, the one-cycle pattern for each process is cropped and transformed into STFT to extract features using both time and frequency. The resulting colormap image is then learned and configured to predict one pattern for a new cycle.

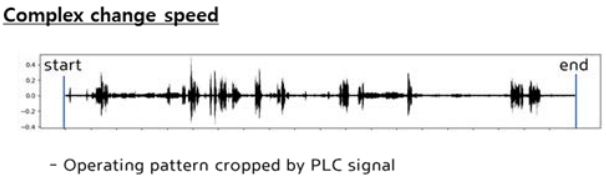


Figure 23. Operation pattern of industrial robot

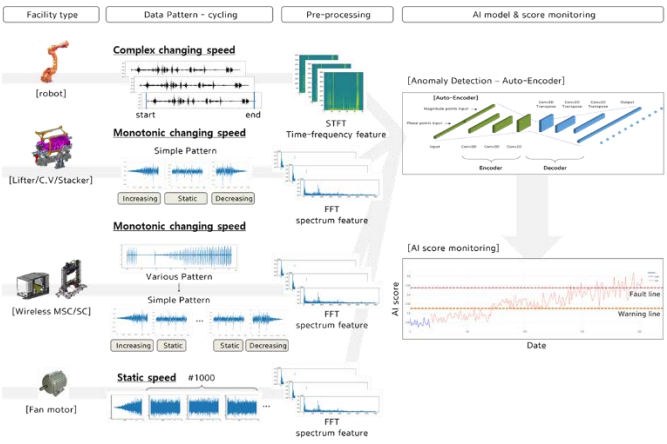


Figure 24. Feature extraction method and anomaly score calculation/monitoring process for various smart factory equipment

The final process is outlined above. When dealing with a fan motor that runs at a constant speed, the frequency spectrum of the continuous speed pattern is used as an input feature. For an engine that drives a stacker/conveyor or wireless mobile device with repetitive acceleration/deceleration, the cycle's acceleration/constant speed/deceleration pattern is separated by the driving part motor and configured through frequency spectrum conversion. For industrial robots with intricate patterns, both time-frequency features are utilized by imaging and

organizing STFT features. An auto-encoder structure defines an input shape for each piece of equipment. The model is saved by learning the data of the normal group. When new data is received, the reconstruction error is calculated and compared against the learned features to determine how much it differs. Finally, an anomaly score is calculated, and the smart factory is monitored by date for anomaly detection.

6.2 Smart Factory operation screen

They are displayed below, as Figure 25 shows the smart factory's configuration screen. The process locations that have undergone PHM algorithm application are marked, allowing for confirmation of AI-predicted vibration data anomalies through Anomaly score monitoring. The score's baseline is partitioned into warning and fault lines. Crossing the warning line triggers a yellow alarm while crossing the fault line triggers a red alarm. Workers can observe the red

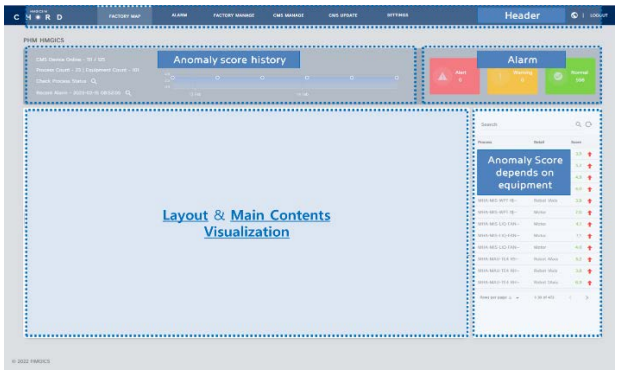


Figure 25. Monitoring screen of the factor layout displayed in the PHM system and each facility's anomaly score

indicator at the factory at the relevant facility location and conduct equipment maintenance activities. The score can be historically tracked by date/time, and a system utilizing MLops has been implemented to enable optimized algorithm re-learning and re-deployment. This system aids in optimizing the anomaly score baseline while operating the factory and conducting intelligent maintenance activities accordingly.

7. Open Problems

Thus far, we have elaborated on developing a PHM system and an abnormality detection framework to effectively address the challenges that may arise when performing predictive maintenance on various facilities within a smart factory. However, there are still some outstanding issues that require attention. Rest assured, we intend to leverage further work to tackle these challenges with precision and efficiency.

7.1. Storage space problem

Despite utilizing edge computing and storing only feature extraction results to address some challenges, many issues still need to be solved. With the emergence of intelligent factories that automate more processes and gather vast amounts of data for analysis, the issue of storage capacity remains a pressing concern.

7.2. The problem of wired sensor installation costs

Vibration data with a high sample rate is ideal for more precise analysis. However, the drawback is that it requires the installation of vibration sensors and the laying of wired cabling. To optimize the return on investment, it is vital to classify equipment based on whether it requires high sample rate data analysis, like vibration.

Installing vibration sensors on every axis would provide valuable data when analyzing robots, but the associated costs would be significant. To mitigate this, it's essential to develop technology that relies on multivariate time series data, such as current, torque, and speed, collected from the robot controller while leveraging multimodal techniques and correlational analysis with vibration to ensure optimal analysis performance without excessive wiring [19], [20].

7.3. Problems classifying various types of defects by line equipment

Efficiently identifying different kinds of defects in manufacturing plants can be challenging. Establishing a seamless collaboration between the operation, analysis, and maintenance departments is crucial. Gathering data on each defect type during factory operations requires meticulous attention to detail and careful feedback collection.

7.4. Problems predicting RUL and lifespan for each line facility

To make precise predictions about facility lifespan [21], data must be collected throughout the entire cycle from the initial operation to the eventual failure. This data can only be obtained once the factory and predictive maintenance system have matured, and the technology can only be developed when the organization and system are well-equipped and consistently engage in seamlessly integrated predictive maintenance activities. With these measures in place, accurate facility lifespan predictions can be confidently made.

7.5. MLOps automation level needs to be increased

The systemization process has been completed; however, users need to re-learn and redistribute the system to utilize its full potential. Once this is done, the system will need further development to enable Continuous Integration and Deployment and subsequently automate MLOps.

7.6. Existing smart factory applications are applied to new facilities.

Due to the implementation of predictive maintenance technology in smart factories, it is anticipated that the emergence of breakdowns caused by aging will be delayed in newly established facilities. As a result, tangible outcomes may take longer to manifest. Given the challenges of collecting defective data in this complex environment, it is imperative to undertake individualized efforts to advance the algorithm.

7.7. Abnormal signal problems, such as simple line failure

In practice, numerous irregular signals may be present alongside the established patterns. These signal

include something as basic as a line fault interruption and can be leveraged in detecting abnormalities due to their distinct data format. In instances with misleading performance data, it is crucial to classify it in a manner that recognizes it as one of the typical states rather than categorizing it as an anomaly.

7.8. Types of failure problems that do not tend to increase gradually

Our team oversees an algorithm designed to identify anomalies by monitoring gradual increases in their score compared to a standard value. However, we recognize that certain types of failures may not exhibit gradual increases, requiring a distinct model for prediction. While we can currently diagnose failures that have already occurred, we strive to advance our technology to enable predictive foresight and prevent these failures altogether.

8. Conclusion

As the demand for predictive maintenance in automated facilities grows alongside the expansion of smart factories, a new study introduces a necessary PHM (Prognostics and Health Management) system and abnormality detection framework method. The system includes a MEMS vibration sensor, Edge CMS device, and PHM platform, while the anomaly detection framework addresses various challenges such as cycling techniques, preprocessing, and AI algorithm development. This methodology effectively addresses data imbalances, learning model limitations, complex equipment patterns, and real-time processing issues commonly faced in manufacturing plants. It also improves upon the problems that arise when deploying such systems in the field, including central server concentration, cycling, classification of various equipment, and algorithm operation problems.

Cutting-edge anomaly detection technology employs an AI-based vibration detection algorithm to collect data at a high sampling rate. It uses edge computing to analyze this data and make real-time decisions. This approach minimizes latency compared to cloud-based analysis and eliminates server capacity issues. The system monitors standard performance data of equipment, learns from it, and provides practical solutions to mitigate issues, ultimately reducing equipment failure and minimizing downtime.

The study has yielded an impressive outcome with the development of abnormality detection technology and PHM systems that are anticipated to enhance the efficiency and sustainability of smart factories. The new smart factory has already achieved mass production, and the challenge of data imbalance in algorithm development has been overcome through data verification. By replacing the reducer of an aging robot in a domestic factory, the technology has proven to be effective. Additionally, the domestic production of these systems can significantly reduce technology investment costs compared to foreign products while allowing for the internalization of HMG's technology. The technology's potential is not limited to smart factories and can be deployed in new facilities such as wireless logistics carriers.

Moving forward, we aim to demonstrate and enhance the efficiency of predictive maintenance in domestic and international automobile factory production line facilities. This vital technology is the backbone of smart and software-defined factories and is poised to assist numerous companies in their pursuit of intelligent automation. Our strategy involves ongoing data monitoring and algorithmic refinement, paving the way for an optimal production line experience.

References

[1] Niklas Tritschler, Andrew Dugenske, Thomas Kurfess. (2021). An Automated Edge Computing-Based Condition Health Monitoring System: With an Application on Rolling Element Bearings. Journal of Manufacturing Science and Engineering. 143(7): 071006 (8ps)

[2] JK Chow, Z Su, J Wu, PS Tan, X Mao. (2020). Anomaly detection of defects on concrete structures with the convolutional autoencoder. Advanced Engineering Informatics. Elsevier, Volume 45, 101105.

[3] Corbinian Nentwich and Gunther Reinhart. (2021). A Method for Health Indicator Evaluation for Condition Monitoring of Industrial Robot Gears. Robotics 2021, 10(2), 80.

[4] M Pech, J Vrchota, J Bednář. (2021). Predictive maintenance and intelligent sensors in smart factory. Sensors, 2021, 21(4), 1470.

[5] Tareq Tayeh, Abdallah Shami. (2021). Anomaly Detection in Smart Manufacturing with an Application Focus on Robotic Finishing Systems: A Review. arXiv:2107.05053 (cs.RO).

[6] A. Bonci, S. Longhi, G. Nabissi and F. Verdini. (2019). Predictive Maintenance System using motor current signal analysis for Industrial Robot. 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Zaragoza, Spain, pp. 1453-1456.

[7] S. K. Bose, B. Kar, M. Roy, P. K. Gopalakrishnan, and A. Basu. (2019). Adepos: Anomaly detection based power saving for predictive maintenance using edge computing. Proceedings of the 24th Asia and South Pacific Design Automation Conference, pp. 597–602.

[8] Pál Péter Hanzelik, Alex Kummer, János Abonyi. (2022). Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development. Sensors 2022, 22(11), 4268.

[9] Ke Feng, J.C. Ji, Qing Ni, Michael Beer. (2023). A review of vibration-based gear wear monitoring and prediction techniques. Mechanical Systems and Signal Processing Volume 182, 109605.

[10] S. S. Patil and J. A. Gaikwad. (2013). Vibration a

of electrical rotating machines using FFT: A method of predictive maintenance. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, pp. 1-6.

[11] George M. Alber and Alan G. Marshall. (1990). Effect of Sampling Rate on Fourier Transform Spectra: Oversampling is Overrated. Appl. Spectrosc. 44, 1111-1116.

[12] Vanraj, Deepam Goyal, Abhineet Saini, S. S. Dhami, B. S. Pabla. (2016). Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques — A review. 2016 International conference on Advances in Computing, Communication, & Automation (ICACCA) (Spring). Dehradun, India, pp. 1-6.

[13] T. Chen, X. Liu, B. Xia, W. Wang and Y. Lai. (2020). Unsupervised Anomaly Detection of Industrial Robots Using Sliding-Window Convolutional Variational Autoencoder. IEEE Access, vol. 8, pp. 47072-47081.

[14] Dan Li, Guoqiang Hu, Costas J. Spanos. (2016). A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis. Energy and Buildings Volume 128, Pages 519-529.

[15] Chaity Banerjee, Tathagata Mukherjee, Eduardo Pasiliao. (2019). An Empirical Study on Generalizations of the ReLU Activation Function. ACM SE '19: Proceedings of the 2019 ACM Southeast Conference Pages 164–167.

[16] Yasi Wang, Hongxun Yao, Sicheng Zhao. (2016). Auto-encoder based dimensionality reduction. Neurocomputing, Volume 184, Pages 232-242.

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna. (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826.

[18] M. Chen, X. Shi, Y. Zhang, D. Wu and M. Guizani. (2021). Deep Feature Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. IEEE Transactions on Big Data, vol. 7, no. 4, pp. 750-758.

[19] Unai Izagirre, Imanol Andonegui, Itziar Landa-Torres, Urko Zurutuza. (2022). A practical and synchronized data acquisition network architecture for industrial robot predictive maintenance in manufacturing assembly lines. Robotics and Computer-Integrated Manufacturing Volume 74, 102287.

[20] Athina Tsanousa, Evangelos Bektsis, Constantine Kyriakopoulos, Ana Gómez González, Urko Leturiondo, Ilias Gialampoukidis, Anastasios Karakostas, Stefanos Vrochidis, Ioannis Kompatsiaris. (2022). A Review of Multisensor Data Fusion Solutions in Smart Manufacturing: Systems and Trends. Sensors 2022, 22(5), 1734.

[21] H. Yan, J. Wan, C. Zhang, S. Tang, Q. Hua and Z. Wang. (2018). Industrial Big Data Analytics for Prediction of Remaining Useful Life Based on Deep Learning. IEEE Access, vol. 6, pp. 17190-17197.

Development of Fault Diagnosis Model based on Semi-supervised Autoencoder

Yongjae Jeon[1], Kyumin Kim[2], YeLim Lee[3], Byeong Kwon Kang[4], and Sang Won Lee[5]

[1,2,3,4] *Department of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
rhrhgudwp@g.skku.edu
aiden7@g.skku.edu
dldpfla1024@g.skku.edu
kevin0730@g.skku.edu

[5] *School of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
sangwonl@skku.edu

ABSTRACT

The maintenance paradigm based on PHM (Prognostics and Health Management) technology, utilizing big data to predict process conditions through manufacturing intelligence, is rising. However, in most industries, there is lack of accurate labeling of sensor data, posing challenges in data utilization due to the significant cost of labeling tasks. Consequently, recent research has focused on semi-supervised learning methodologies, which are applicable in label-absent scenarios. Especially, there is a growing emphasis on semi-supervised autoencoder, which learns both labeled and unlabeled data simultaneously. Also, there is a demand for the development of fault diagnosis models for essential components, such as bearings in most mechanical systems. Vibrational data is actively being integrated with artificial intelligence for application in bearing fault diagnosis frameworks. Nonetheless, diagnosing the condition of bearings inside machine systems, especially within the machine tool spindle, remains challenging, as the labeling of collected data causes significant costs. Therefore, this paper aims to develop a fault diagnosis model for unlabeled bearings in machine tool spindle using a semi-supervised autoencoder. Initially, a monitoring system of bearing simulator that imitates a machine tool spindle bearing was constructed, and vibration signals from both normal and fault bearings were collected based on this system. Subsequently, a semi-supervised autoencoder model was developed to construct a fault diagnosis model using labeled simulator data and unlabeled machine tool spindle bearing data. To evaluate the model, additional data of normal and fault bearings in machine tool spindle were collected, and the performance of the model was compared with a conventional fault diagnosis model based on 1D-CNN.

1. INTRODUCTION

In manufacturing, machine system faults not only degrade the quality of the products but also lead to downtime, resulting in significant costs. Therefore, Prognostics and Health Management (PHM) techniques utilizing sensor data and artificial intelligence are widely used to monitor equipment and diagnose failures. In particular, deep learning methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are used to enhance the reliability of failure diagnosis (Wen, Li, Gao & Zhang, 2017; Ince, Kiranyaz, Eren, Askar & Gabbouj, 2016; Cabrera, Guamán, Zhang, Cerrada, Sanchez, Cevallos, Long & Li, 2020; Abed, Sharma, Sutton & Motwani, 2015). However, acquiring a large amount of high-quality data for deep learning training is challenging in industries. While data are collected in various processes, almost data lack labels due to the high cost of labeling. Therefore, although data are abundant, distinguishing between normal and faulty states is difficult, posing challenges for developing failure diagnosis models. Particularly bearings, the core components of machine tools such as spindles, disturb the stable operation of the spindle when they break down because real-time confirmation and labeling of bearing failure are difficult.

Therefore, this paper proposes the utilization of data from a similar domain system and a semi-supervised autoencoder model to diagnose faults in unlabeled bearings of machining spindle bearings. First, a testbed capable of simulating the rotational motion of machining spindle bearings is constructed to create an environment for collecting enough labeled bearing data. Then a semi-supervised autoencoder

structure, which can learn unlabeled real-world data and labeled testbed data, is used for fault diagnosis model. By learning unlabeled data, domain difference between the actual equipment and the testbed is reduced, allowing for the extraction of generalized features and thus the development of a fault diagnosis model applicable to actual equipment. To verify its effectiveness, the performance of the proposed model is compared with that of a conventional model trained only on testbed data.

This paper introduces the proposed model and its applications in Chapter 2. Chapter 3 describes the data collection process for machining spindle bearings and the testbed. Modeling is performed in Chapter 4, and a comparison of the performance between the conventional model and the proposed model is also conducted.

2. METHODS

2.1. Autoencoder

Autoencoder is one of the unsupervised learning algorithms, that has an Encoder neural network that reduces the dimension of input data, and a Decoder neural network that reconstructs the input data from the reduced dimensions. Both networks are connected through the latent space, to reconstruct the original data from the latent space where features of the data are preserved. Typically, the Mean Squared Error (MSE) is used as the loss function to minimize the difference between the input data and the reconstructed data. Hinton and Salakhutdinov (2006) confirmed Autoencoder is available for data dimensionality reduction, and Kingma and Welling (2013) proposed the Variational Autoencoder structure, which combined with probabilistic models, for data generation applications. Additionally, Sakurada and Yairi (2014) confirmed its usability in anomaly detection.

2.2. Semi-supervised Autoencoder

Semi-supervised learning is a method of training model where one part of the input data is labeled (Reddy, Viswanath & Reddy, 2018). This approach is typically used in situations where labeled and unlabeled data are mixed. Semi-supervised learning uses labeled data to train the model and unlabeled data to improve the model's generalization performance. This offers the advantage that utilizing data efficiently and building models with a shortage of labeled data.

Semi-supervised Autoencoder is the conventional Autoencoder structure with a separate Fully-Connected layer(FC layer) to enable learning from labeled data. For the task of classification, a Classifier can be added to the Encoder and Decoder structure as shown in Figure 1. In this structure, labeled data is used to train the Classifier to perform well in classification from the reduced dimensions by the Encoder. Also, with the unlabeled data the Autoencoder is trained to extract generalized features. Additionally, the loss function is defined as a joint loss combined with the cross-entropy for classification and the Mean Squared Error(MSE) for reconstruction, as shown in Eq. (1). Each weight is one of the hyperparameters that need to be optimized during the training process. As a result, by using both labeled and unlabeled data in training, an Encoder that extracts generalized features and a Classifier that has high classification performance can be obtained.

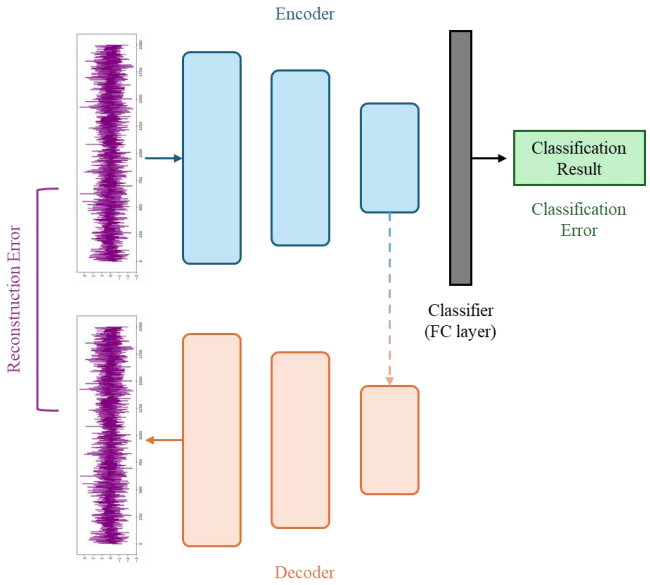


Figure 1. Semi-supervised Autoencoder structure

$$L_J = w_R L_R + w_C L_C \quad (1)$$

2.3. Proposed Method

In this paper, labeled bearing simulator(source) data, which is similar to an unlabeled machining spindle bearing(target) data, is used for fault diagnosis using the Semi-supervised Autoencoder, as shown in Figure 2. Initially, both labeled source data and unlabeled target data are used to train a feature extractor(Encoder) and reconstructor(Decoder). This allows the feature extractor to extract generalized features that reduce domain difference between target and source. Furthermore, the labeled source data is additionally used to train the feature extractor and classifier, making the classifier determine a decision boundary that can classify normal and fault conditions from generalized features. Through this structure and method, it can diagnose normal and fault conditions of the unlabeled target data.

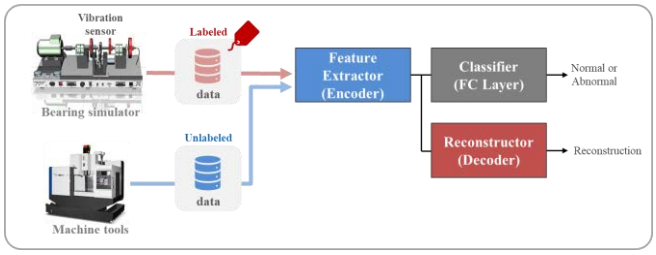


Figure 2. Diagram of proposed method

3. Experiment and Data Acquisition

3.1. Machining spindle bearing

The machining spindle bearing, which is the target domain for fault diagnosis, has a 6204 ball bearing inside the spindle, as illustrated in Figure 3. An accelerometer is attached to the spindle for data acquisition. While the spindle was rotating at 2,000 RPM, 200 data were collected every 0.1 seconds with 20,000Hz sampling frequency without overlapping, so each data had 2,000 points. However, this data was collected without labels, lacking information about normal or fault conditions. Therefore, to develop a fault diagnosis model, it is necessary to use a source domain that is similar to the machining spindle bearing, but with information on the condition.

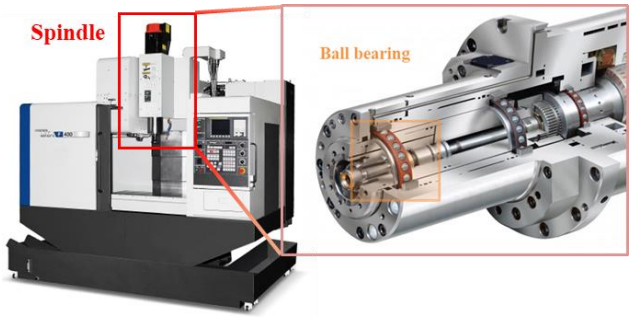


Figure 3. Image of machining spindle bearing

3.2. Bearing Simulator

The bearing simulator is a source domain to imitate the operation of a machining spindle bearing, as shown in Figure 4. It connects the motor on the left and the bearing on the right by the shaft to allow rotation. The attached bearing is the 6204 ball bearing used in the machining spindle, and an accelerometer is installed in the Y direction to collect data during rotation. The way of data acquisition was identical to those for the machining spindle bearing, and 400 data are each collected using both normal and fault bearings.

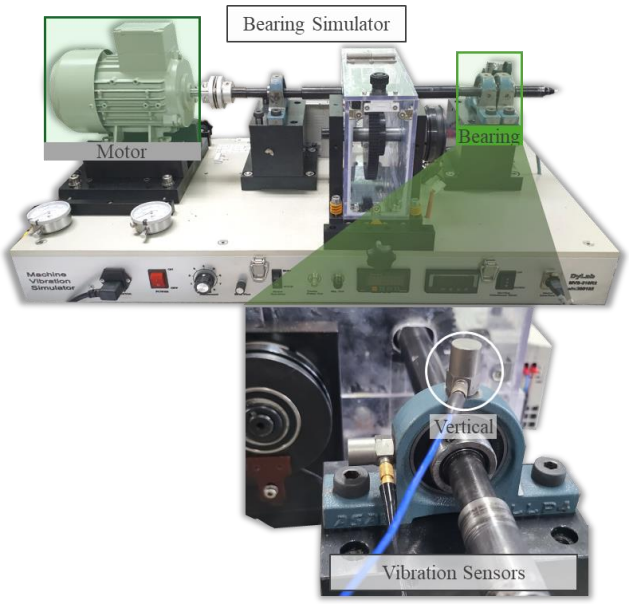


Figure 4. Image of bearing simulator

4. Modeling

4.1. Model Architecture

The architecture of the proposed model is explained in Table 1. The Feature Extractor and Reconstructor were composed of 1D convolutional layers and transposed 1D convolutional layers, and the hidden layer of the classifier was set as one, as shown in Figure 5. A 1D convolutional layer is a network that replaces vertical and horizontal convolution with unidirectional convolution to apply convolution operation in vector data. Through this layer, time-series data can be used in training without converting into a matrix.

Table 1. Description of proposed model architecture

Network	Layer type
Feature Extractor (Encoder)	Conv. layer 1
	Conv. layer 2
	Max Pooling layer
	Conv. layer 3
	Conv. layer 4
Reconstructor (Decoder)	Transposed Conv. layer 1
	Transposed Conv. layer 2
	Up sampling
	Transposed Conv. layer 3
	Transposed Conv. layer 4
Classifier (FC)	Hidden layer
	Output layer

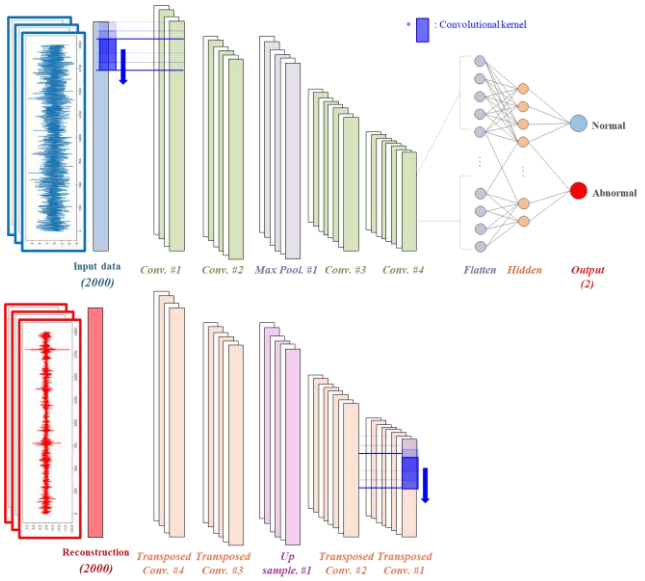


Figure 5. Proposed model architecture

4.2. Model Training

While training the model, the hyperparameters were set as follows. The number of filters in all convolutional layers was 16, the neuron of the hidden layer was set to 16, kernel size was set to 7, and stride was set to 5. ReLU was used for the activation function for all layers, except the output layer, which was set to softmax. The optimizer was Adam, with a learning rate of 0.0001, the weights for classification error (w_C) and reconstruction error (w_R) in the joint loss were set to 0.5 arbitrarily. A total of 700 iterations were done for training with a mix of 200 unlabeled target domain data and 800 labeled source domain data.

To evaluate the performance of the proposed model, a conventional fault diagnosis model was additionally trained. It had only a feature extractor and classifier as shown in Figure 6, with the same model structure and hyperparameters as the proposed model. Since this model can only be trained with labeled data, it is trained with only 800 labeled source domain data.

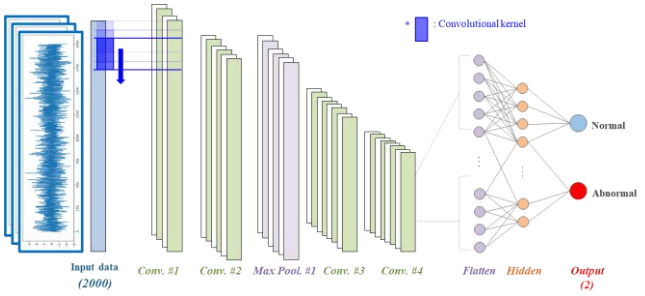
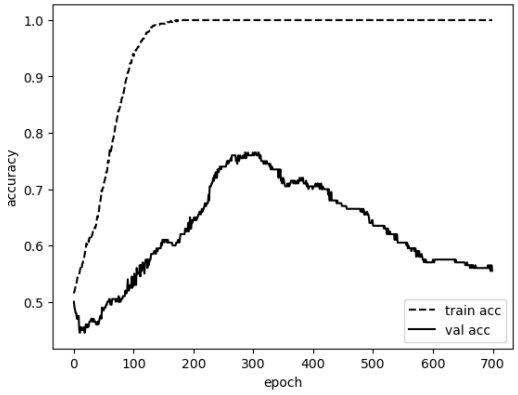


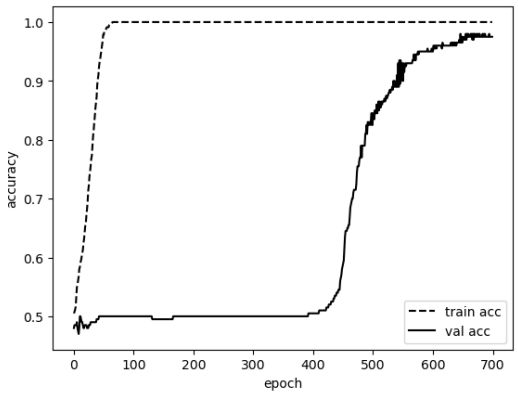
Figure 6. Conventional model architecture

4.3. Result

Model performance comparison was done using additional validation data, collected by machining spindle bearing. Normal and fault data were collected by operating the machining spindle at 2,000 RPM, with normal bearings and bearings damaged by impacts. Under the same setting as the training data acquisition, 100 data for each normal and fault condition were collected. Using these data, the accuracy of the proposed model and conventional model were evaluated to compare the performance of the models. The final accuracy was 76.5% for the conventional model and 97.5% for the proposed model which the proposed model has higher performance. Figure 7 shows the accuracy of both train and validation data at each epoch for both models. Both models reached 100% accuracy on the train data. However, the validation accuracy of the conventional model initially shows an increase, but the accuracy decreases in the end part because of the overfitting problem caused by domain differences. In contrast, the proposed model's validation accuracy did not increase significantly at the beginning but reached high accuracy at the end. This happens due to the delay in optimizing the autoencoder's error compared to the classifier, which can be improved by changing the weight ratio.



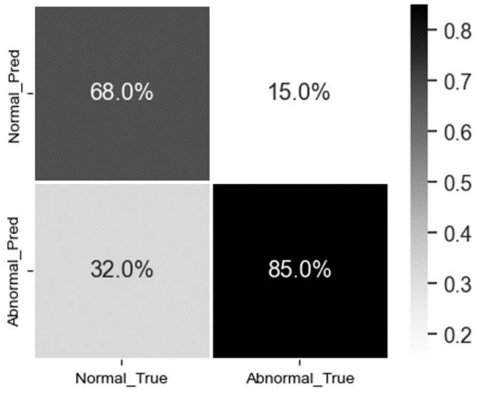
(a) Conventional model



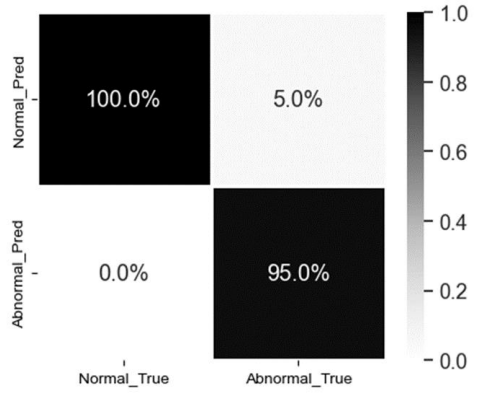
(b) Proposed model

Figure 7. Training and validation accuracy at each epoch

Figure 8 is a confusion matrix that compares the accuracy of the proposed and conventional models with validation data. Only 5% of the fault data are misdiagnosed by the proposed model, whereas 15% of fault data are misdiagnosed by the conventional model. Also, the conventional model misdiagnoses 32% of normal data, resulting in low performance.



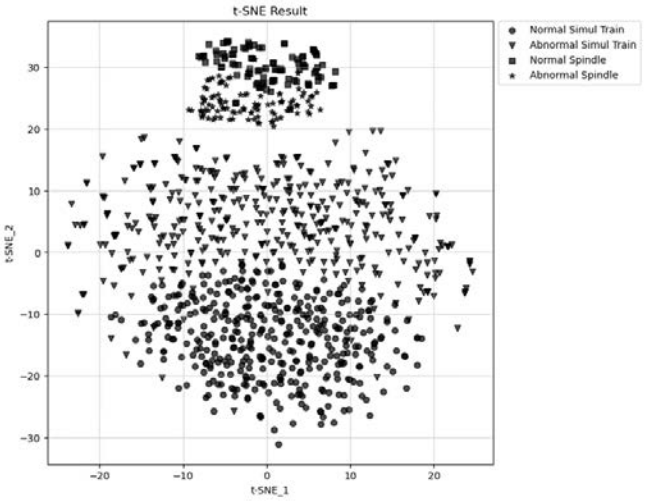
(a) Conventional model



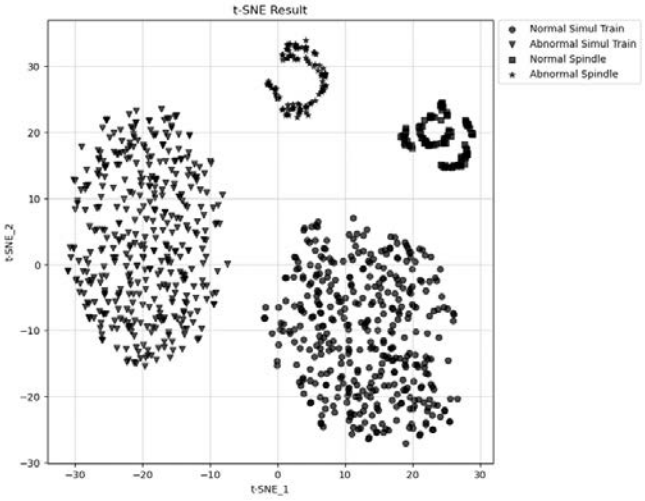
(b) Proposed model

Figure 8. Confusion matrix of final accuracy

Figure 9 visualizes features extracted by the feature extractor from the source domain data used for training and the target domain data used for validation, using t-distributed Stochastic Neighbor Embedding (t-SNE). Unlike the conventional model, which makes it hard to define a decision boundary due to the domain differences, the proposed model can make a clear decision boundary, showing reduced domain differences.



(a) Conventional model



(b) Proposed model

Figure 9. Result of t-SNE by feature

5. CONCLUSION

This paper proposes a fault diagnosis methodology with multi-domain data using a semi-supervised autoencoder to solve the problem of developing fault diagnosis models with the lack of label data in the real-world industry. A testbed, similar to the domain of actual equipment, was constructed to train the proposed model with sufficient labeled data and unlabeled equipment data. From this approach, a feature extractor that extracts generalized features by reducing the influence of domain information, and a classifier that can diagnose conditions based on generalized features were developed. The model was validated with additional machining spindle bearing data, resulting in the development of a high-performance fault diagnosis model. This approach enables the practical utilization of unlabeled data collected

from industrial machines. Furthermore, it has been demonstrated that a high-performance fault diagnosis model can be developed with unlabeled data. This can be applied to many manufacturing, contributing to the reduction of labeling costs across various industries.

However, this study conducted a fault diagnosis model from the perspective of deep learning without considering physical method. In case of bearings, the amplitude of certain fault frequency increase depending on fault characteristics of bearings. The reason for not applying this method is that the value at the fault frequency does not appear clearly, because of the various signals of the machine tool. Therefore, just monitoring the value in the fault frequency is not appropriate for bearing fault diagnosis of the machine tool. However, by considering both data-driven model and physical method, a hybrid model with higher performance can be developed.

And this study did not include the uncertainty caused by various operating conditions of bearings, as it focused on a diagnostic model under fixed operating conditions. Also, the comparison with other models, that can use the unlabeled data, was insufficient. Therefore, the development of a fault diagnosis model that can robustly operate under changing bearing operating conditions and validation of the effectiveness of the proposed methodology through comparison with other semi-supervised learning, synthetic data generation techniques, and domain adaptation technologies are planned for future works.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (No. 2022R1A2C3012900) and the Ministry of Trade, Industrial and Energy (No. 20012807) grant funded by the Korea government.

NOMENCLATURE

L_J	Joint loss
w_R	Reconstruction weight
L_R	Reconstruction loss
w_C	Classification weight
L_C	Classification loss

REFERENCES

Wen, L., Li, X., Gao, L., & Zhang, Y. (2017). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, 65(7), 5990-5998. doi: 10.1109/TIE.2017.2774777

Ince, T., Kiranyaz, S., Eren, L., Askar, M., & Gabbouj, M. (2016). Real-time motor fault detection by 1-D convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 63(11), 7067-7075. doi: 10.1109/TIE.2016.2582729

Cabrera, D., Guamán, A., Zhang, S., Cerrada, M., Sanchez, R. V., Cevallos, J., Long, J. & Li, C. (2020). Bayesian approach and time series dimensionality reduction to LSTM-based model-building for fault diagnosis of a reciprocating compressor. *Neurocomputing*, 380, 51-66. doi: https://doi.org/10.1016/j.neucom.2019.11.006

Abed, W., Sharma, S., Sutton, R., & Motwani, A. (2015). A robust bearing fault detection and diagnosis technique for brushless DC motors under non-stationary operating conditions. Journal of Control, *Automation and Electrical Systems*, 26, 241-254. doi: https://doi.org/10.1007/s40313-015-0173-7

Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Walle, R. V. & Hoecke, S. V. (2016). Convolutional neural network based fault detection for rotating machinery. *Journal of Sound and Vibration*, 377, 331-345. doi: https://doi.org/10.1016/j.jsv.2016.05.027

Lu, C., Wang, Z., & Zhou, B. (2017). Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification. *Advanced Engineering Informatics*, 32, 139-151. doi: https://doi.org/10.1016/j.aei.2017.02.005

Jiang, H., Li, X., Shao, H., & Zhao, K. (2018). Intelligent fault diagnosis of rolling bearings using an improved deep recurrent neural network. *Measurement Science and Technology*, 29(6), 065107. doi: 10.1088/1361-6501/aab945

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504-507. doi: 10.1126/science.1127647

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. doi: https://doi.org/10.48550/arXiv.1312.6114

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. *In Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis* (pp. 4-11). doi: https://doi.org/10.1145/2689746.2689747

Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: A brief review. *International Journal of Engineering &Technology,* 7(1.8) (pp. 81-85). doi: https://doi.org/10.14419/ijet.v7i1.8.9977

BIOGRAPHIES

Yongjae Jeon is now Ph.D. candidate in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree (System Management Engineering) from Sungkyunkwan University in 2020. His research interest is AI-based prognostics and health management for smart manufacturing.

Kyumin Kim is now master candidate in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree (Mechanical Engineering) from Sungkyunkwan University in 2023. His research interest is AI-based prognostics and health management for smart manufacturing.

Yelim Lee is now master candidate in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. She obtained her bachelor's degree (Mechanical Engineering) from Sungkyunkwan University in 2023. Her research interest is AI-based prognostics and health management for smart manufacturing.

Byeong Kwon Kang has a master's degree in the Sustainable Design and Manufacturing Laboratory, Department of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree (Mechanical Engineering) from Sungkyunkwan University in 2022 and completed his master's degree in 2024. His research interest is AI-based prognostics and health management for smart manufacturing.

Sang Won Lee is now professor in the school of Mechanical Engineering, Sungkyunkwan University. He obtained his bachelor's degree in 1995 and master's degree in 1997 (Mechanical Design and Production Engineering) from Seoul National University. He obtained his Ph.D. degree (Mechanical Engineering) from University of Michigan in 2004. His research interest includes prognostics and health management (PHM), cyber-physical system (CPS), additive manufacturing, and data-driven design.

DiffPhysiNet: A Bearing Diagnostic Framework Based on Physics-Driven Diffusion Network for Unseen Working Conditions

Zhibin Guo[1] Jingsong Xie[2] Tongyang Pan[3] and Tiantian Wang[4]

[1,2,3,4] *School of Traffic & Transportation Engineering, Central South University, China*

marcogzb@csu.edu.cn
jingsongxie@foxmail.com
ty.pan@csu.edu.cn
wangtt@hnu.edu.cn

ABSTRACT

Fault diagnosis is essential to ensure bearing safety in industrial applications. Many existing diagnostic methods require large scales of data from a full range of working conditions. However, the structure and working conditions differences between machines lead to significant variation in data distribution, making it difficult to diagnostic with unseen samples. To handle this situation, an unknown condition diagnosis Framework (UCDF) based on physics-driven diffusion network (DiffPhysiNet) is proposed, effectively integrating the generation capability of the diffusion model and learning from the working conditional encoding (WCE). Specifically, signals under limited working conditions are gradually convert to noise through a forward noising process. Then, DiffPhysiNet reconstructs signals from the noise by a reverse denoising process. In addition, a physics-driven UNet (Physi-UNet) structure is designed to extract WCE for noise level prediction during the reverse process. Moreover, an Unsupervised Clustering Filter (UCFilter) is constructed to select signals with high quality after generation. Signals under unknown working condition can be generated with certain WCE. Ultimately, extensive experiments on two bearing datasets (SDUST and PU) validate the effectiveness of our method compared with the state-of-the-art baselines and the ablution test confirms the significant role of Physi-UNet and UCFilter.

1. INTRODUCTION

Rotating machinery is crucial in modern industry, highlighting the need for effective condition monitoring and fault diagnosis technology to ensure its security and reliability (Kordestani et al. 2021). Deep learning-based approaches have gained significant attention in machine condition monitoring as a data-driven fault diagnosis method(Zio 2022).

For deep learning to effectively diagnose faults in rotating machinery, it requires consistency in the data distribution between training and testing sets. However, practical industrial applications often present challenges that hinder the applicability of deep learning methods, which can be concluded as follows: **(1)** Rotating components often operate under varied conditions, such as changes in rotational speed and load (Chen and Li 2017). **(2)** Obtaining sufficient labeled data with precise health information across all operating conditions can be impractical. **(3)** Domain shift issues arise when attempting to compensate for information gaps by utilizing labeled data from multiple machines or different working conditions. Also, due to discrepancies in data (Ben-David et al. 2010).

In recent years, various advanced techniques have been developed to tackle the aforementioned challenges. One of these techniques is domain adaptation (DA), which aims to reduce the distribution discrepancy between the source and target domains during model training (Wang et al. 2020). DA allows the transfer of knowledge acquired from large source datasets to construct diagnostic models for smaller target datasets with similar characteristics. Wang et al. proposed the use of intra-class maximum mean discrepancy (MMD) in conjunction with multi-scale ResNet architectures to reduce the conditional distribution discrepancy of vibration signals (Wang et al. 2020). Hu et al. introduced tensor-aligned invariant subspace learning, which enables the discovery of a shared tensor representation for cross-domain diagnosis cases (Hu, Wang, and Gu 2020). Inspired by adversarial learning principles, Li et al. developed a method to map knowledge from target to source working conditions using generative adversarial networks (Li et al. 2021). Domain adaptation techniques can improve the robustness and generalization capabilities of fault diagnosis models. However, these methods are limited by the closed-set assumption, meaning that the source and

target domains have feature distributions that cannot be crossed (Si et al. 2021).

Under this premise, it is necessary to develop a technique that takes the out-of-distributed (OOD) fault classification into account (Michau and Fink 2019). Generative Adversarial Networks (GANs) nowadays adopts an unsupervised learning method and automatically learns from the source data. In the applications of PHM, conditional GANs have been used to control the generation process to generate desired distinct classes. Wang et al. introduced an enhanced version of Least-Square Generative Adversarial Networks (LSGANs) which notably retain more signal details compared to traditional methods and exhibit significant robustness (Wang et al. 2019). However, these methods are only suitable for generating data from previously observed conditions and not for generating previously unseen conditions in a specific domain (Rombach, Michau, and Fink 2023). The latter is the focus of our research. For the issue of the diagnostic works under unseen working conditions, propose a new framework for Open-Partial DA based on generating distinct fault signatures with a Wasserstein GAN, which enables a better transferability between two different domains (Li et al. 2022). However, the main drawback of GANs is that they are unstable during the training process and it is hard to embed diagnostic knowledge during the process of generation(Cui et al. 2023). Nowadays, same as a generative model, diffusion model does not suffer from GANs-like problems of training non-convergence and pattern collapse.

To achieve the stable and effective diagnostic framework for unseen working conditions based on feature embedding, we first propose a Physics-Driven Diffusion Network (DiffPhysiNet) for unknown condition diagnosis, which can generate a complete bearing sample of industrial environments and maintain the real-world working conditions through physics-informed methods. DiffPhysiNet effectively integrates the generation capability of the diffusion model and embeds working conditional encoding (WCE). Essentially, the forged signals generated by DiffPhysiNet guarantee the generation accuracy while retaining the utility. To summarize, the primary contributions of this work are concluded as follows:

- A denoising diffusion-based generative model DiffPhysiNet is proposed, which can generate high-quality signal data.
- A novel neural network structure called Physi-UNet, which integrates the residual block and attention mechanism to model signal features of bearings.
- UCFilter is constructed based on K-means clustering method to select the valuable signals after generation.

The remainder of the paper is structured as follows: In Section 2, we provide background information relevant to our research and in Section 3, we formally introduce the proposed diagnostic framework and its components in detail. Section 4 outlines the dataset utilized and presents the experimental results and in Section 6, we summary our work and proposes some research directions for the future.

2. PRELIMINARY

In this section, we first briefly introduce the basic knowledge of Denoising Diffusion Probabilistic Models (DDPM) (Shu, Li, and Farimani 2023) and Fourier Neural Operator (FNO)(Rafiq, Rafiq, and Choi 2022), which are the fundamentals of the proposed DiffPhyisNet.

2.1. Diffusion model

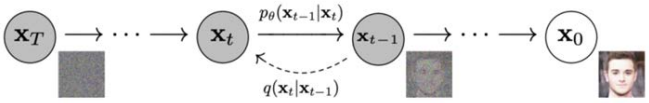


Figure 1. Two processes in Denoising Diffusion Probabilistic Models

As shown in Figure **1**, the diffusion model that typically contains two processes: forward process and reverse process. In this setting, a sample from the data distribution $x_0 \sim q(x)$ is gradually noised into a standard Gaussian noise $x_T \sim \mathcal{N}(0, I)$ by the forward process, where the transition is parameterized by $q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right)$ with $\beta_t \in (0,1)$ as the amount of noise added at diffusion step t.

A neural network learns the reverse process of gradually denoising the sample via reverse transition $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\, \mu_\theta(x_t,t), \sum \theta(x_t,t))$ .Learning to clean x_T through the reversed diffusion process can be reduced to learning to build a surrogate approximator to parameterize $\mu_\theta(x_t,t)$ for all t. The denoising model $\mu_\theta(x_t,t)$ can be trained by using a weighted mean squared error loss which we will refer to as:

$$\mathcal{L}(x_0) = \sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \|\mu(x_t,x_0) - \mu_\theta(x_t,t)\|^2 \qquad (1)$$

where $\mu(x_t,x_0)$ is the mean of the posterior $q(x_t \mid x_{t-1})$. This objective can be justified as optimizing a weighted variational lower bound on the data log likelihood. Also note that the original parameterization of $\mu_\theta(x_t,t)$ can be modified in favor of $\hat{x}_0(x_t,t,\theta)$or$\epsilon_\theta(x_t,t)$.

2.2. Fourier neural operator

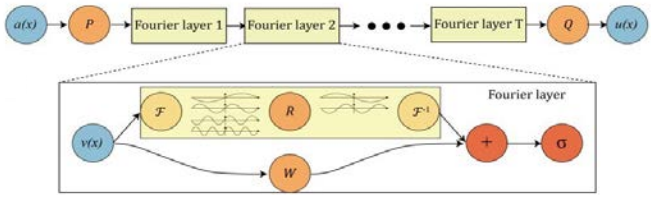


Figure 2. The full architecture of neural operator

The main idea of FNO is to use Fourier transform to map high-dimensional data into the frequency domain and approximate nonlinear operators by learning the relationships between Fourier coefficients through neural networks. The FNO architecture is shown in **Figure 2**, which consists of three main steps:

a) The input $a(x)$ is lifted to a higher dimensional representation $v_0(x) = P\big(a(x)\big)$ by the local transformation P, which is commonly parameterized by a shallow fully connected neural network.

b) The higher dimensional representation $v_0(x)$ is updated iteratively by:

$$v_{t+1}(x) = \sigma\big(W v_t(x) + (\mathcal{K}(a;\phi)v_t)(x)\big) \qquad (2)$$

where $(\mathcal{K}(a;\phi)v_t)(x)$ is a linear transform on the frequency domain of the amplitude and the phase of $v_t(x)$, W: is a linear transform on the high-dimension of the time domain. σ: is the non-linear activation function.

c) The output $u(x)$ is obtained by $u(x) = Q\big(v_T(x)\big)$, where Q : is the projection of v_T , and it is parameterized by a fully connected layer.

$\mathcal{F}$ and $\mathcal{F}^{-1}$are denoted as Fourier transform and its inverse transform of a function, allowing the operations on the frequency domain of the high-dimension. The Fourier neural operator (FNO)(Lehmann et al. 2024) aims to map between two infinite-dimensional spaces by training on a finite set of input–output pairs. It has been demonstrated that the FNO can serve as a universal approximator capable of accurately representing any continuous operator.

3. PROPOSED METHOD

In this section, we elaborate on the proposed DiffPhysiNet framework as shown in Figure 3. We start by presenting the diagnostic principles and steps of the proposed method under unseen working conditions. Then, we introduce the details of the denoising model, i.e., Physi-UNet. Furthermore, UCFiler utilize K-means to cluster some generated sample to prove the generation quality is also introduced.

3.1. Diagnostic principles of DiffPhysiNet

As illustrated in **Figure 3**, aiming at the diagnostic under unseen working conditions, 4 parts (a-d) are involved in the framework.

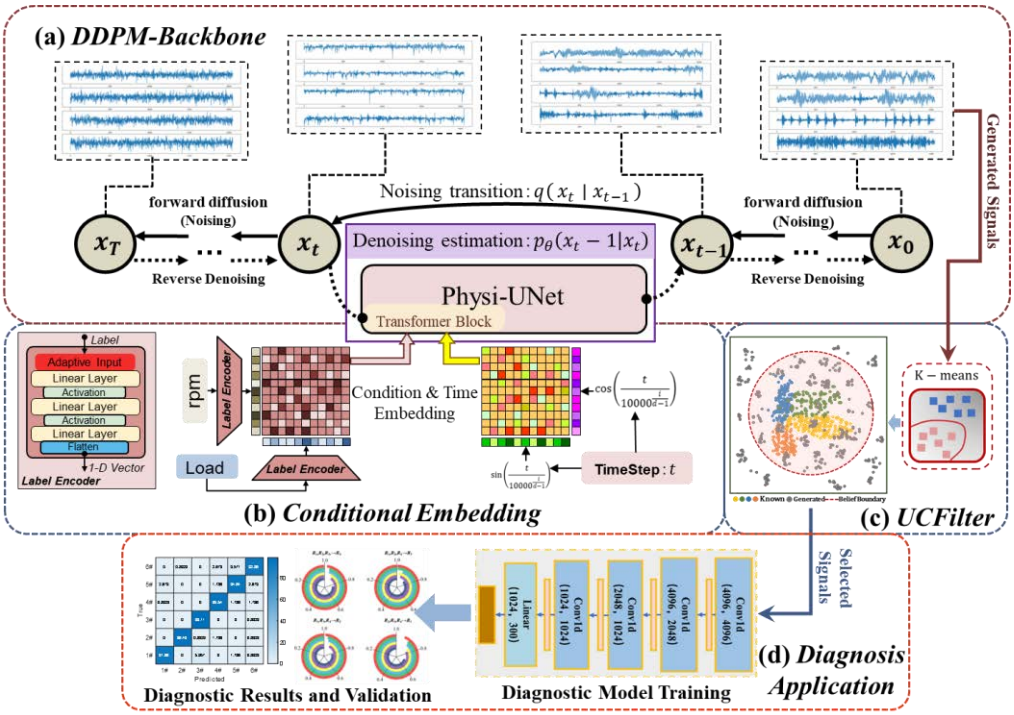


Figure 3. The diagram of proposed DiffPhysiNet framework

(a) The first part is based on the diffusion model which is introduced in **2.1**, utilizing this generative model rather than other generative methods mainly attribute to the style embedding convenience which ensures that generated fault signatures contain physics-driven features. (b) The second part aims to construct a latent space with conditional

encoding methods, making sure of that the projection space of working conditions are continuous. (c) The third part utilizes an unsupervised clustering method to select qualified generated signals guaranteeing the effectiveness of the training datasets. (d) The last part is the application stage of this proposed method, utilizing the selection of generated signals for the diagnostic model training. Then the validated diagnostic model can be applied for online fault diagnosis.

3.2. Structure of Physi-UNet

As aforementioned, the diffusion model (DDPM), A neural network learns the reverse process of gradually denoising the sample via reverse transition $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\theta(x_t,t), \sum\theta(x_t,t))$ .Learning to clean x_T through the reversed diffusion process can be reduced to learning to build a surrogate approximator to parameterize $\mu_\theta(x_t,t)$ for all t. In our proposed framework, Physi-UNet is utilized for the process of denoising estimation, of which the structure is shown in **Figure 4**. As illustrated in the figure, an implicit U-Net is introduced to enhance Fourier neural operator. The denoised signal after $(T-t)$ steps is utilized as the input, which is then converted into a high-dimensional representation via the lifting layer P, and finally the output is obtained through the projection of Q, converting the vectors from a high-dimensional space to 1D vibration signal (Benitez et al. 2023) .

The structural design of the Physi-UNet is based on the hypothesis that the Fourier spectrum of fault data can be expressed as the sum of (1) domain-specific components (the spectrum of a signal from normal operation) and (2) of fault-specific components representing the specific fault characteristics. In other words, this hypothesis allows us to express Fourier coefficients (Dang and Ishii 2022) of the fault data of a certain class c from a specific domain $\mathbb{X}$ ($x^{c,FFT}_{fault,\mathbb{X}}$) as a sum of domain-specific characteristics that are represented by the domain features $x^{FFT}_{\mathbb{S}}$ and the fault class specific characteristics that are domain-independent $x^{c,FFT}_{fault}$ $fault$ and scaled by a factor w, which can be expressed by:

$$x^{c,FFT}_{fault,\mathbb{X}} = x^{FFT}_{\mathbb{X}} + w * x^{c,FFT}_{fault} \tag{3}$$

The physics-driven fault component and the domain specific features are demodulated based on the embedding of Times Step t and the continuous working conditional encoding (WCE), which is of great importance on the guidance of feature decomposition. As shown in **Eq. (4)**.

$$\begin{cases} Attention(Q,K,V) = softma\,\boldsymbol{x_t}\left(\dfrac{QK^\top}{\sqrt{d}}\right)\cdot V \\ x^{c,FFT}_{fault,\mathbb{X}} = \boldsymbol{x_t} + Attention(Q,K,V) \\ Q = W_Q\cdot x_t, K = W_K\cdot x_t, V = W_V\cdot x_t \end{cases} \tag{4}$$

where $\boldsymbol{x_t} \in \mathbb{R}^{c\times n}$ (c and n represent the dimensions and length of the signal) is the given input, W_Q, W_K, and W_V are learnable parameter matrices from the embedded Timestep t and the encoded working condition.

Hence, the fault components of the vibration signals are captured in using Fourier bases:

$$A^{(k)}_{i,t} = \left|\mathcal{F}(x^{c,FFT}_{fault,\mathbb{X}})_k\right|, \Phi^{(k)}_{i,t} = \phi\left(\mathcal{F}(x^{c,FFT}_{fault,\mathbb{X}})_k\right), \tag{5}$$

$$\kappa_{i,t}{}^{(1)},\cdots,\kappa^{(K)}_{i,t} = \underset{k\in\left\{1,\cdots,\left\lfloor\frac{\tau}{2}\right\rfloor+1\right\}}{arg\ TopK}\left\{A^{(k)}_{i,t}\right\}, \tag{6}$$

$$P_{i,t}(x) = \sum_{k=1}^{K} A^{\kappa^{(k)}_{i,t}}_{i,t}\ cos\left(2\pi f_{\kappa^{(k)}_{i,t}}\tau c + \Phi^{\kappa^{(k)}_{i,t}}_{i,t}\right), \tag{7}$$

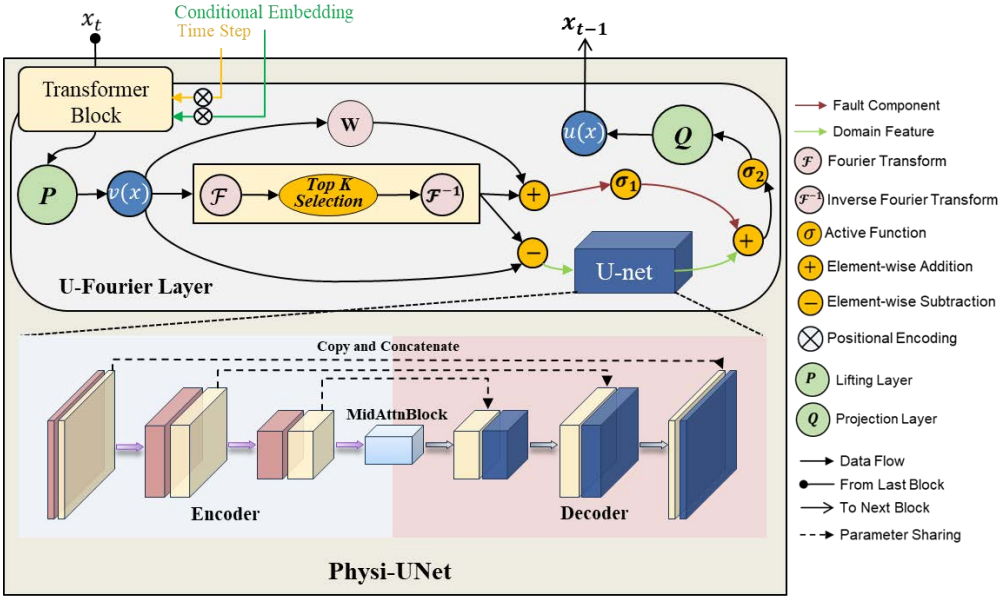


Figure 4. The structure of proposed Physi-UNet enhanced with FNO and U-Net component

where arg TopK is to get the top K amplitudes and K is a hyperparameter. $A_{i,t}^{(k)}$, $\Phi_{i,t}^{(k)}$ are the phase, amplitude of the k-th frequency after the discrete Fourier transform $\mathcal{F}$ respectively. f_k represents the Fourier frequency of the corresponding index k. In fact, the Fourier layer selects bases with the most significant amplitudes in the frequency domain, and then returns to the time domain through an inverse transform to model the physics-driven fault features.

The U-Net structure is utilized to synthesis the corresponding domain features according to the residue component after removing the fault frequencies.

$$D_{i,t}(x) = UNet(v_0(x) - P_{i,t}(x)) \tag{8}$$

where the $P_{i,t}$ is selected fault component and the $D_{i,t}$ is the obtained domain features.

$D_{i,t}$ and $P_{i,t}$ are then reweighted and activated in the following process, the summation and projection combine the physic-driven and domain feature components, which can be expressed as:

$$C_{i,t}(x) = \sigma_2\left[UNet(v_0(x) - P_{i,t}) + \sigma_1[W(P_{i,t})]\right] \tag{9}$$

$$x_{t-1} = Q\left(C_{i,t}(x)\right) \tag{10}$$

where $\sigma_1[\cdot]$ and $\sigma_2[\cdot]$ are the activation function, $W(\cdot)$is the reweight layer, $Q(\cdot)$ is the projection layer aforementioned, $x_{t-1}$is the out put of this diffusion step.

3.3. Unsupervised Clustering Filter

Once the DDPM in the DiffPhysiNet training is completed, the working conditions of signal generation is controlled by the WCE. As shown in Figure **5**. the generated signal clustered by K-means algorithm, which is an unsupervised clustering method. The top of n samples nearest to the signal sample center selected as valuable signals.

The distances between inter-class samples are measured by distribution probability based on Kullback-Leibler divergence between the joint probabilities R_{ij} in the high-dimensional space and the joint probabilities T_{ij} in the low-dimensional space. The values of R_{ij} are defined to be the symmetrized conditional probabilities, whereas the values of T_{ij} are obtained by means of the Student's t-distribution with one degree of freedom. The calculation is summarized as follows:

$$r_{ij} = \frac{r_{i|j} + r_{j|i}}{2n} \tag{11}$$

$$t_{ij} = \frac{\left(1 + |y_i - y_j|^2\right)^{-1}}{\sum_{k \neq l}(1 + |y_k - y_l|^2)^{-1}} \tag{12}$$

where $r_{i|j}$ is the distribution probability of sample point j when the sample point i is given. y is the generated samples. The values of r_{ii} and t_{ii} are set to zero. The calculation of the Kullback–Leibler divergence C_d between the two joint probability distributions R and T is given as follows:

$$C_d = KL(R\,||T) = \sum_i \sum_j r_{ij} \log r_{ij} - r_{ij} \log t_{ij} \tag{12}$$

After calculated the Kullback–Leibler divergence C_d of every sample in generated signals, the selection boundaries are decided according to the expectation numbers of acceptable samples n, and k value of the ***Boundary Decision*** indicates the portion of selected samples (k=n/N).

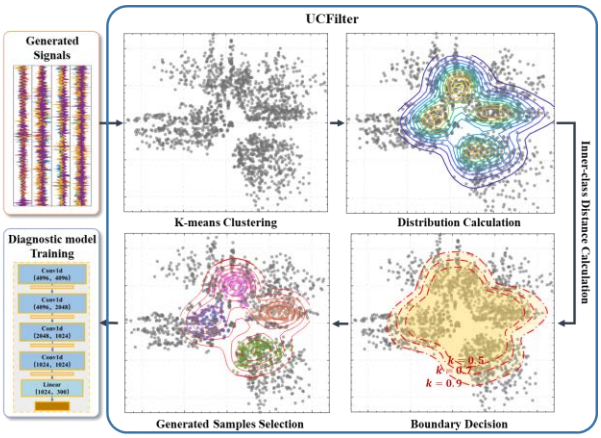


Figure 5. The steps of unsupervised clustering filter method

4. Experimental Validation

4.1. Experiments Setting

To assess the diagnostic model's performance in unfamiliar conditions, we conducted two experimental case studies using test rigs from SDUST (Jia et al. 2020) and the Paderborn University bearing dataset (PU dataset) for bearing fault diagnosis. The first case involves a constant domain shift with the rig operating at various constant speeds, while the second case examines a constant domain shift with the rig operating under multiple conditions of variable rotational speeds and loads. These experiments confirm the effectiveness of the DiffPhysiNet method, which utilizes diffusion models, for complex industrial applications.

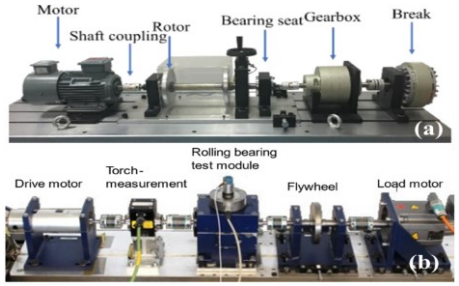


Figure 6. Experimental platform of SDUST dataset (a) and PU dataset (b).

4.1.1. Description of SDUST dataset and Case1

Figure **6** (a) shows the experimental platform of SDUST, which includes a motor, a shaft coupling, a rotor, a testing bearing, a gearbox, and a break.The bearing type utilized is N205EU, with data collected across four health conditions: normal (NOR), inner ring fault (I), rolling element fault (B), and outer ring fault (O). Four distinct working conditions were tested at speeds of 1000, 1500, 2000, and 2500r/min. The experiment sets four diagnostic cases for unseen working conditions across domains: T1000, T1500, T2000, and T2500 as shown in Table 1.

Table 1. **Case1:** Diagnostic cases of SDUST dataset.

Diagnostic cases	Seen Domain	Unseen Domain
T1	1500r/min,2000r/min,2500r/min	1000r/min
T2	1000r/min,2000r/min,2500r/min	1500r/min
T3	1000r/min,1500r/min,2500r/min	2000r/min
T4	1000r/min,1500r/min,2000r/min	2500r/min

4.1.2. Description of PU dataset and Case2

The test rig of PU dataset is shown in Figure **6** (b), which is mainly composed of a motor, a torque measurement shaft, a bearing test module, a flywheel, and a load motor. There are 7 health conditions, normal (N), inner-race fault (IF) with three damage levels (IF1, IF2, and IF3), outer-race fault (OF) with two damage levels (OF1 and OF2), and compound fault (CF) containing IF and OF.

Faulty bearings with real damage were acquired from an accelerated lifetime test. Vibration data was collected under four distinct working conditions, involving rotational frequency (Hz), load torque (Nm), and radial force (N), at a sampling frequency of 64 kHz. These conditions create four domains: P1, P2, P3, and P4, leading to four diagnosis cases, as outlined in Table 2. Each category in unseen working conditions comprises 2000 samples.

Table 2. **Case2:** Diagnostic cases of PU dataset.

Domain	Rotational frequency	Load	Diagnostic Cases	Seen Domain	Unseen Domain
p1	25Hz	1000N	**R1**	*p2, p3, p4*	*p1*
p2	15Hz	400N	**R2**	*p1, p3, p4*	*p2*
p3	15Hz	1000N	**R3**	*p1, p2, p4*	*p3*
p4	25Hz	400N	**R4**	*p1, p2, p3*	*p4*

4.1.3. Compared methods

Some typical or up-to-date technologies were utilized as a set of compared methods to validated the effectiveness of the DiffPhysiNet framework with the idea of Physi-UNet and UCFilter and all the methods used the same preprocessing and network back-bone for a fair comparison. As shown in Table 3, M1-M6 series are competitive related methods, M1 means the Domain Adaption (DA) method based on empirical risk minimization (ERM) principle using multi-domain data based on the general cross-entropy loss of DA method. M2-M4 follow the same setting in (Jiao et al. 2020; Huang et al. 2022; Han, Li, and Qian 2021) by adding a distance metric or distribution alignment as a loss term, such as MMD, JMMD, and CORAL. M5 (Li et al. 2020) is a start-of-the-art method DG that uses adversarial training with normalization strategies and a strategy of multi-case training and M6 (Chen et al. 2022) is a competitive method for cross-domain diagnosis under unseen domain through triplet loss and data augmentation with Gaussian noise.

Table 3. Related methods for comparation

Methods	Description
M1	DA
M2	DA with MMD
M3	DA with JMMD
M4	DA with CORAL
M5	ADIG (Li et al. 2020)
M6	IEDGNet (Chen et al. 2022)

4.2. Experimental results and analysis

4.2.1. Generated signal and analysis

To assess the effectiveness of the proposed diagnostic framework, we illustrate the generated signals in Figure **7**. Notably, there were no fault samples with a 1000N load and an rpm of 40Hz in the training set, yet similar fault samples were generated. Thus, DiffPhysiNet can generate fault signals for unseen combinations of rpm and load.

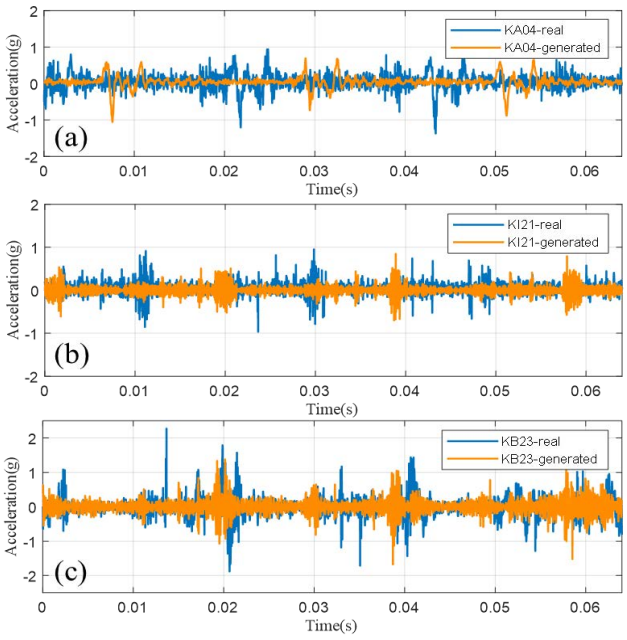


Figure 7. Generated signals comparison of time-domain on Set *R1* of *PU dataset* and real signals (a)Inner ring fault (b) Outer ring fault (c) Cage fault.

Comparing the time-domain of the generated signals with real signals reveals a high degree of similarity, confirming the ability of the proposed method to generate signals under unseen working conditions that closely resemble real signals. This verifies the effectiveness and practicality of the generative model. On the frequency-spectrum comparison between the generated signals and the real signals is shown in **Figure 8**. The frequency spectrum of the generated signal and the real signal has a high degree of coincidence, especially in the low-frequency band where contains most damage features according to the vibration theory of bearing fault.

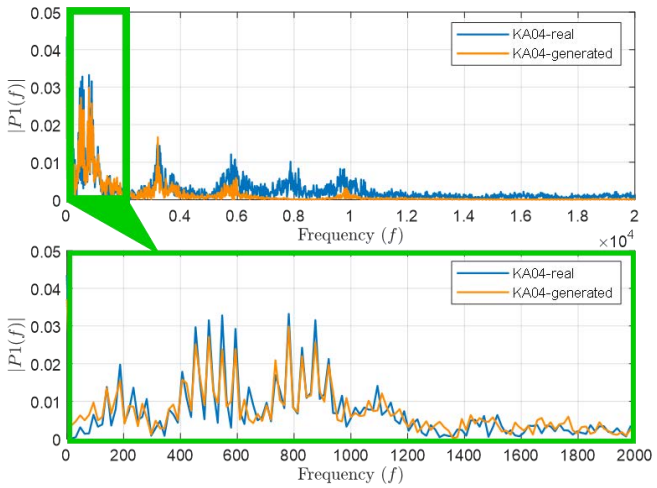


Figure 8. Generated signals comparison of frequency-domain on Set *R1* of *PU dataset* and real signals (a)Inner ring fault (b) Outer ring fault (c) Cage fault.

4.2.2. Experimental results and analysis

Table 4. presents the diagnostic results for the proposed method and comparison methods in Case 1 (SDUST dataset) and Case 2 (PU dataset). Several conclusions can be drawn. Firstly, the basic DA method achieves notably lower average test accuracies of 76.46% and 80.63% in the two cases respectively, indicating interference among data distributions during training, affecting model generalization. Secondly, DA methods using MMD, JMMD, and CORAL exhibit improvements over basic ERM, with average accuracy gains of 1.57%, 12.06%, and 10.05% respectively in Case 1, and 4.04%, 5.27%, and 3.49% in Case 2. These methods aim to eliminate distributional discrepancies between source domains and learn domain-invariant representations. Finally, the proposed method achieves the best diagnostic performance in almost all DA cases, with the highest average accuracy of 94.15% and 93.27% in both experimental cases. Furthermore, it demonstrates superior stability compared to com-parison methods in most cases. **Figure 9.** illustrates the classification accuracy of different diagnostic cases for Case 1 and Case 2, aiding in the comparison of diagnostic results.

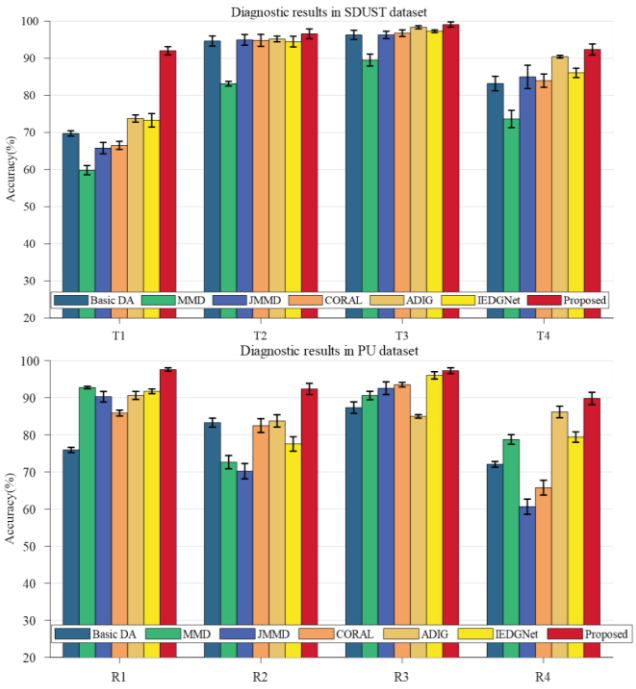


Figure 9. Classification accuracy of the different diagnostic cases of case 1 and case 2.

Table 4. Diagnostic accuracy (%) of Case 1 and Case2.

	M1	M2	M3	M4	M5	M6	Proposed
Case1							
T1	59.73±1.43	59.85±2.48	65.8±3.05	66.51±2.2	73.75±1.91	73.27±3.64	93.03±2.17
T2	84.65±2.72	83.14±1.28	94.95±2.9	94.81±3.27	95.19±1.52	94.48±2.86	95.6±2.64
T3	86.28±2.47	89.51±3.21	96.28±1.98	96.76±1.8	98.35±0.84	97.27±0.74	98.06±1.42
T4	83.19±3.89	83.62±4.67	84.98±6.26	83.96±3.6	90.42±0.76	86.06±2.49	91.37±2.94
Average	76.46±2.63	78.03±4.66	88.52±3.55	86.51±2.72	88.42±1.26	88.77±2.93	94.15±2.54
Case 2							
R1	75.92±1.38	92.75±0.69	90.26±2.87	85.86±1.59	90.61±2.17	91.73±1.26	97.6±0.92
R2	83.26±2.42	72.64±3.57	70.19±4.15	82.47±3.73	83.74±3.33	77.54±3.89	92.36±3.02
R3	87.33±3.07	90.59±2.26	92.57±3.42	93.52±1.24	84.98±0.94	96.02±1.96	97.29±1.55
R4	72.04±1.53	78.73±2.59	60.61±4.11	65.73±3.97	86.17±3.12	79.39±2.78	89.81±3.31
Average	80.63±2.10	84.67±2.27	79.41±3.64	85.89±2.63	87.38±2.39	87.17±2.47	93.27±2.21

4.2.3. Feature visualization and analysis

Compared methods based on domain adaptation (DA) aim to learn domain-invariant features, while our diagnosis seeks to generate physic-embedded signals to cover unseen distributions. we employ feature visualization to validate these conclusions. To illustrate the distribution of fault features from seen and unseen domains corresponding to seen and unseen working conditions, we present 2-D features from the second layer of the fault classifier using T-SNE (van der Maaten and Hinton 2008). For clarity, we plot the feature vectors of the SDUST dataset (Case 1) from four health conditions under case T1and four colors represent four domains, with gray points indicating features extracted from unseen domains, while other colors denote features from available source do-mains.

The domain adaptation (DA) based method aims to extract generalized features that are consistent across different domains, including unseen domains. However, as illustrated in **Figure 10**, the features learned by methods M1-M3 fail to capture the generalized representation of the I02 fault due to significant domain discrepancies between seen and unseen working conditions. Although method M4 may learn more robust features compared to M1-M3, it struggles to cluster effectively in the seen domains of I02.

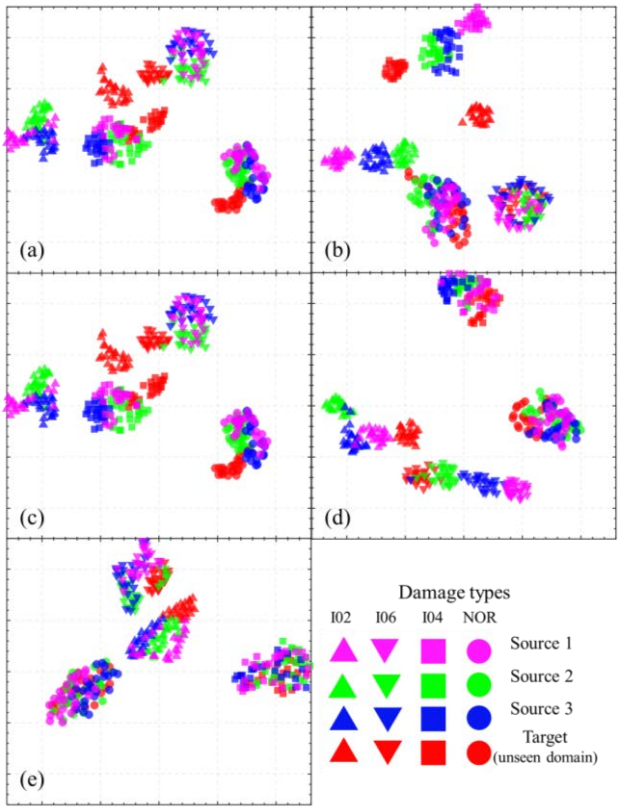


Figure 10. Results of feature-dimension reduction via T-SNE under unseen target working condition:M1, (b) M2, (c) M3, (d) M4, (e) Proposed.

In contrast, the proposed DiffPhysiNet method, leveraging the feature embedding capability of Physi-UNet, can generate signals with more domain-invariant features. This leads to improved classifier training and reaffirms that the Physi-UNet structure enables the model to fit not only the source data distribution but also data from unseen working conditions.

4.2.4. Parameter sensitivity analysis

Adjustable parameters are involved in the construction and training of the proposed method and considering their impact on the model performance, parameter sensitivity analyses are performed on all case.

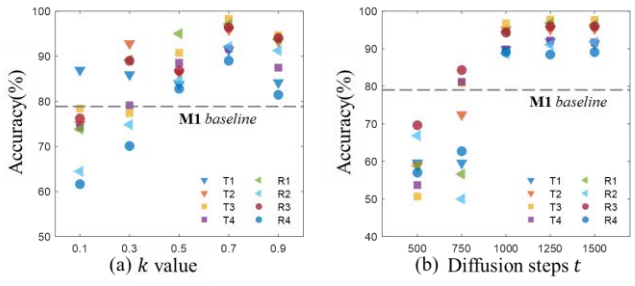


Figure 11. Parameter sensitivity analysis (a) k value: the portion of selected nearest generated samples. (b) Diffusion steps t: the iteration steps of generation progress.

k value represents the proportion of selected samples ($k = n / N$), where N is the total number of generated samples and n is the number of selected samples among them. These selected samples undergo the nearest distribution process in the K-means clustering algorithm, as depicted in **Figure 5**, where the Boundary Decision is determined by the k value. Figure **11**(a) illustrates the diagnostic accuracy for each case under different k values ranging from 0.1 to 0.9. The graph indicates that when k is below 0.5, meaning less than half of the total generated samples are selected, the performance is inferior to the baseline M1 method. This could be attributed to the reduced generation capability resulting from a smaller set of selected samples, causing offsets in the feature distribution from the real distribution. Conversely, if k is too large, the model's performance declines, likely due to the utilization of too many substandard generated signals for diagnostic model training. Hence, selecting a value of k around 0.7 is recommended.

t is the parameter of diffusion steps, the larger the value of diffusion steps the more detailed the signal generation process will be, as well as the larger the model training time and the computational resources it will consume. As shown in Figure **11** (b), it can be concluded that the performance of DiffPhysiNet framework becomes better and more stable as t is bigger than 1000 steps. Regarding the performance of M1 method as the baseline, the diffusion steps should not be less than 1000. As t increased from 1000 to 1500, the

performance of the model stabilizes and does not improve significantly.

5. CONCLUSION

In conclusion, we propose the DiffPhysiNet framework for diagnosing bearing faults under unseen working conditions for safety-critical equipment. Leveraging a generative diffusion model and working conditional encoding (WCE), this framework effectively embeds signal features, and the UCFilter method ensures signal quality using principles from K-means clustering. Experimental validation on real-world bearing datasets demonstrates the superiority of Physi-UNet over existing approaches, particularly in diagnostic accuracy. Feature visualization confirms the framework's ability to capture generalized signal features under unknown conditions, highlighting the generative model's efficacy in signal generation.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 52302446) and National Natural Science Foundation of China Excellent Youth Fund (Grant No. 52322215).

REFERENCES

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. 2010. 'A theory of learning from different domains', *Machine Learning*, 79: 151-75.

Benitez, Jose Antonio Lara, Takashi Furuya, Florian Faucher, Xavier Tricoche, and Maarten V. de Hoop. 2023. 'Fine-tuning Neural-Operator architectures for training and generalization', *Arxiv*.

Chen, L., Q. Li, C. Q. Shen, J. Zhu, D. Wang, and M. Xia. 2022. 'Adversarial Domain-Invariant Generalization: A Generic Domain-Regressive Framework for Bearing Fault Diagnosis Under Unseen Conditions', *Ieee Transactions on Industrial Informatics*, 18: 1790-800.

Chen, Z. Y., and W. H. Li. 2017. 'Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network', *Ieee Transactions on Instrumentation and Measurement*, 66: 1693-702.

Cui, W., J. Ding, G. Y. Meng, Z. Y. Lv, Y. H. Feng, A. M. Wang, and X. W. Wan. 2023. 'Fault Diagnosis of Rolling Bearings in Primary Mine Fans under Sample Imbalance Conditions', *Entropy*, 25.

Dang, Z. R., and M. Ishii. 2022. 'Towards stochastic modeling for two-phase flow interfacial area predictions: A physics-informed reinforcement learning approach', *International Journal of Heat and Mass Transfer*, 192.

Han, T., Y. F. Li, and M. Qian. 2021. 'A Hybrid Generalization Network for Intelligent Fault Diagnosis of Rotating Machinery Under Unseen Working Conditions', *Ieee Transactions on Instrumentation and Measurement*, 70.

Hu, C. F., Y. X. Wang, and J. W. Gu. 2020. 'Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks', *Knowledge-Based Systems*, 209.

Huang, Z. L., Z. H. Lei, G. R. Wen, X. Huang, H. X. Zhou, R. Q. Yan, and X. F. Chen. 2022. 'A Multisource Dense Adaptation Adversarial Network for Fault Diagnosis of Machinery', *Ieee Transactions on Industrial Electronics*, 69: 6298-307.

Jia, S. X., J. R. Wang, B. K. Han, G. W. Zhang, X. Y. Wang, and J. T. He. 2020. 'A Novel Transfer Learning Method for Fault Diagnosis Using Maximum Classifier Discrepancy With Marginal Probability Distribution Adaptation', *Ieee Access*, 8: 71475-85.

Jiao, J. Y., M. Zhao, J. Lin, and K. X. Liang. 2020. 'Residual joint adaptation adversarial network for intelligent transfer fault diagnosis', *Mechanical Systems and Signal Processing*, 145.

Kordestani, M., M. Saif, M. E. Orchard, R. Razavi-Far, and K. Khorasani. 2021. 'Failure Prognosis and Applications-A Survey of Recent Literature', *Ieee Transactions on Reliability*, 70: 728-48.

Lehmann, F., F. Gatti, M. Bertin, and D. Clouteau. 2024. '3D elastic wave propagation with a Factorized Fourier Neural Operator (F-FNO)', *Computer Methods in Applied Mechanics and Engineering*, 420.

Li, R. R., S. M. Li, K. Xu, J. T. Lu, G. R. Teng, and J. Du. 2021. 'Deep domain adaptation with adversarial idea and coral alignment for transfer fault diagnosis of rolling bearing', *Measurement Science and Technology*, 32.

Li, W. H., R. Y. Huang, J. P. Li, Y. X. Liao, Z. Y. Chen, G. L. He, R. Q. Yan, and K. Gryllias. 2022. 'A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges', *Mechanical Systems and Signal Processing*, 167.

Li, X., W. Zhang, Q. Ding, and J. Q. Sun. 2020. 'Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation', *Journal of Intelligent Manufacturing*, 31: 433-52.

Michau, G., and O. Fink. 2019. 'Domain Adaptation for One-Class Classification: Monitoring the Health of Critical Systems Under Limited Information', *International Journal of Prognostics and Health Management*, 10.

Rafiq, M., G. Rafiq, and G. S. Choi. 2022. 'DSFA-PINN: Deep Spectral Feature Aggregation Physics

Informed Neural Network', *Ieee Access*, 10: 22247-59.

Rombach, K., G. Michau, and O. Fink. 2023. 'Controlled generation of unseen faults for Partial and Open-Partial domain adaptation', *Reliability Engineering & System Safety*, 230.

Shu, D. L., Z. J. Li, and A. B. Farimani. 2023. 'A physics-informed diffusion model for high-fidelity flow field reconstruction', *Journal of Computational Physics*, 478.

Si, Y. N., J. X. Pu, S. F. Zang, and L. F. Sun. 2021. 'Extreme Learning Machine Based on Maximum Weighted Mean Discrepancy for Unsupervised Domain Adaptation', *Ieee Access*, 9: 2283-93.

van der Maaten, L., and G. Hinton. 2008. 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9: 2579-605.

Wang, K. W., X. Zhang, Q. S. Hao, Y. Wang, and Y. Shen. 2019. 'Application of improved least-square generative adversarial networks for rail crack detection by AE technique', *Neurocomputing*, 332: 236-48.

Wang, X., C. Q. Shen, M. Xia, D. Wang, J. Zhu, and Z. K. Zhu. 2020. 'Multi-scale deep intra-class transfer learning for bearing fault diagnosis', *Reliability Engineering & System Safety*, 202.

Zio, E. 2022. 'Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice', *Reliability Engineering & System Safety*, 218.

Zuo, L., F. J. Xu, C. H. Zhang, T. F. Xiahou, and Y. Liu. 2022. 'A multi-layer spiking neural network-based approach to bearing fault diagnosis', *Reliability Engineering & System Safety*, 225.

BIOGRAPHIES

Zhibin Guo was born in Zhengzhou, Henan, China, in 2001. He is currently pursuing the D.E. degree in transportation engineering with Central South University, Changsha, China. His current research interests include rotating machine diagnosis, PHM for railway infrastructure, health management of highspeed train bogie, and deep learning.

Jingsong Xie was born in Anren, Hunan, China, in 1989. He received the B.S. degree from the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China, in 2013, and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, in 2018. He joined the School of Traffic and Transportation Engineering, Central South University, Changsha, China, as a Lecturer. His research interests include fault diagnosis, machine learning, vibration analysis, and crack diagnosis.

Tongyang Pan was born in Shaanxi, China, in 1994. He received the B.S. and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2017 and 2022 respectively. He is currently a lecturer with Central South University, Changsha, China. His research interests include mechanical signal processing, intelligent fault diagnosis, and machinery condition monitoring

Tiantian Wang received the bachelor's and Ph.D. degrees from Beihang University, Beijing, China, in 2012 and 2018, respectively. He is currently a Professor with Central South University, Changsha, China, and Hunan University, Changsha. His current research interests include vehicle aerodynamics and vehicle structure, especially train/tunnel aerodynamics, and prognostics and health management (PHM) for trains.

Domain Adaptation for Fault Detection in Civil Nuclear Plants

Henry Wood[1], Felipe Montana [2], Visakan Kadirkamanathan [3], Andy Mills [4], Will Jacobs[5],

[1, 2, 3, 4] *The Department of Automatic Control and Systems Engineering*
The University of Sheffield
Amy Johnson Building, Portobello Street, Sheffield, S1 3JD, United Kingdom
henry.wood@sheffield.ac.uk
f.montana-gonzalez@sheffield.ac.uk
visakan@sheffield.ac.uk
a.r.mills@sheffield.ac.uk
w.jacobs@sheffield.ac.uk

ABSTRACT

Recent domain adaptation approaches have been shown to generalise well between distant data domains achieving high performance in machine fault detection through time series classification. An interesting aspect of this transfer-learning inspired approach, is that the algorithm need not be exposed to fault data from the target domain during training. This promotes the application of these methods to environments in which fault data is unfeasible to obtain, such as the detection of loss-of-coolant accidents (LOCA) in nuclear power plants (NPPs).

A LOCA is a failure mode of a nuclear reactor in which coolant is lost due to a physical break in the primary coolant circuit. If undetected, or not managed effectively, a LOCA can result in reactor core damage.

Three high-fidelity physics based models were created with divergent behaviour that represent different data domains. The first model is used to generate source domain data by simulating labelled training data under both nominal and LOCA conditions. The second and third models act as surrogates of real plants and are used to generate target domain data, i.e. to simulate nominal data for training and LOCA condition data for validation.

Several deep-learning feature encoders (with varying levels of connectivity) were applied to this LOCA detection problem. Among these, a 'Baseline' encoder was used to quantify the improvement that domain adaptation techniques make to LOCA detection performance under large domain divergences.

Classification accuracy for each model is explored within the context of LOCA break size and location within each plant model.

The proposed method for LOCA detection demonstrates how the dependence upon sparse accident-specific data can be alleviated through the use of domain adaptation. Detection capability of the LOCA condition is maintained even when no data examples are available in the target domain.

1. INTRODUCTION

There is an opportunity in the nuclear industry to adopt data-driven methods to help maintain the safe operation of critical systems, both as a result of the improved availability of sensing instrumentation and the rapid evolution of network architectures for fault detection (Gomez-Fernandez et al., 2020). A plethora of approaches for identifying abnormal transient behaviour, such as a LOCA, exist with foundations in probabilistic methods (Aldemir, 2013), complex fluid-structure interaction models (Mahmoodi et al., 2011) and Markov modelling (Sakurahara et al., 2019). Although effective in finite environs, conventional methods suffer when attempting to compensate for the lack of available labelled fault data in the nuclear domain.

A LOCA occurs when a physical break in the reactor coolant system releases coolant faster than recovery systems can replenish it. This increases the temperature of the core, which can damage the plant and potentially release reactivity. Detection of this transient behaviour is paramount to the safe operation of pressurised water reactors (PWRs).

Time series classification through deep learning methods has seen increased attention (Ismail Fawaz et al., 2019) in previous years, with myriad techniques being derived to tackle a spectrum of fault detection problems (Wei & Keogh, 2006). The nuclear industry has received its share of attention in this regard, with neural network based methodologies tasked with

aiding the monitoring of numerous aspects of NPPs, including diagnosing the source of abnormalities in operation data (Lee et al., 2021) and tuning a digital twin to provide supplementary NPP data (Wang et al., 2021).

Existing examples of these methods simultaneously identify and characterise transients whilst making remaining useful life predictions (Rivas et al., 2024). Typically, though, these approaches rely upon the assumptions that similar quantities of nominal and faulty data exist, and that data gathered from differing NPP sources (physical plants and simulations alike) will share a similar data distribution.

One branch of deep-learning research aims at tackling this manner of problem through Transfer Learning. Current works in industrial contexts display impressive results regarding fault diagnosis with minimal labelled training data under diverse application domains (Y. Zhang et al., 2023), as well as combinations of global and local models providing more robust remaining useful life predictions (J. Zhang et al., 2023). Domain adaptation (a subset of Transfer learning where data sources share the same input space) can simultaneously make data gathered from multiple sources appear more similar, whilst separating sub-classes within those sources, eg. 'Normal' and 'Faulty' data (Qian et al., 2023).

Existing LOCA detection procedures that attempt to overcome the issue of the lack of available NPP accident data perform well in a limited range of operating conditions (Farber & Cole, 2020). The generalised knowledge available through leveraging transfer-learning from attainable model data has not yet been fully exploited in the context of LOCA detection. Domain adaptation allows the transfer of the knowledge contained in such a classifier on to a new domain containing previously unseen behaviour.

In this work, we introduce adaptations to current transfer learning based fault detection methods with application to the detection of the LOCA condition. The model design process was guided by system experts in order to construct data 'features' that well represent the NPP behaviour in both nominal and LOCA conditions. Results show how domain adaptation is able to retain the fault detection performance that is achievable for the labelled training data when it is applied to the surrogate models data domain.

2. PROBLEM FORMULATION

2.1. Domain adaptation overview

Consider data sampled from two distinct domains, Source (S) and Target (T). The data from each domain (x_S and x_T, respectively) posses different distributions. Additionally, suppose that class labels for the data sampled from the Target domain, y_T, are unavailable. Given the data is drawn from disparate distributions, conventional supervised methods cannot infer knowledge about the Target domain using data from the Source domain.

Domain adaptation provides methods for prediction of target domain labels y_T from target domain data x_T using the information present in source domain data and labels x_S and y_S. In this work, we will describe a feature extractor as an encoder: a network designed to construct a feature space Z using the distributions of x_S and x_T. The aim of the encoder is to provide a transformation through which the distributions of x_S and x_T appear similar to each other in the feature space Z.

The generalised feature space Z is used to aid classification for samples from the target domain, since the encoded representations of x_S and x_T are similar, and we have access to class labels for the source domain data, y_S. There exist many well documented methods by which the encoder can construct Z, with two of the most commonly used domain-adaptation specific measures being Mutual Information (MI) and Maximum Mean Discrepancy (MMD).

2.1.1. Mutual Information

MI is a statistical quantity that describes how much information one variable conveys about another. If we consider these variables in terms of the feature space representations of the input domain data, i.e: z_S and z_T (obtained from passing x_S and x_T respectively into the transformative encoder), then maximising the MI between the Target domain feature space representation (z_T) and the entire feature space (Z) will encourage the encoder to generate features that are generalised between the two input domains.

The MI between these specific variables can be expressed as a linear combination of the Shannon Entropy of each feature space representation (Chen et al., 2021). The Shannon Entropy for a distribution A is given by

$$H(A) = -\sum_{a \in A} P(a) \ln P(a). \tag{1}$$

If we state that Z_S and Z_T are the distributions of the feature space representations z_S and z_T respectively, then the MI becomes

$$\text{MI}(Z_T; Z) = -\sum_{z_S \in Z_S} P(z_S) \ln P(z_S) - \sum_{z_T \in Z_T} P(z_T) \ln P(z_T). \tag{2}$$

Maximising this quantity during training promotes the generation of features that convey the largest amount of shared information between the Target domain samples and the entire set of observed samples from each domain.

2.1.2. Maximum Mean Discrepancy

A brief description of the intended function of the MMD term will be sufficient for understanding its relevance to this work. The key principal that underpins MMD metrics is the idea that if two distributions are equal, then their statistical properties should also be equal. By using MMD, it is possible to perform a hypothesis test upon the functions that transform the input domain distributions into their encoded feature representations. These functions are embedded as a Hilbert space, a convenient mathematical construct which allows linear algebra to be applied to infinite-dimensional vectors.

Formally, the MMD between two distributions A and B on the sets X and Y can be calculated as

$$\begin{aligned}\text{MMD}(A,B) &= ||\mathbb{E}_{X\sim A}[\phi(X)] - \mathbb{E}_{Y\sim B}[\phi(Y)]||_H \\ &= \sup_{f\in H}(\mathbb{E}_{X\sim A}[f(X)] - \mathbb{E}_{Y\sim B}[f(Y)]), \quad (3)\end{aligned}$$

where f is a function in the Hilbert space H and ϕ is the transformation from the input set to the Hilbert space. The supremum means this is equivalent to taking the maximum of the mean difference between the distributions A and B.

In practice, the mean of the feature-space distributions is not known, so the MMD between the two feature-space distributions must be empirically estimated by

$$\begin{aligned}\text{MMD}(Z_S,Z_T) = &\frac{1}{m(m-1)}\sum_{i}^{m}\sum_{j\neq i}^{m}\phi(z_{S_i},z_{S_j}) \\ &- 2\frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n}\phi(z_{S_i},z_{T_j}) \\ &+ \frac{1}{n(n-1)}\sum_{i}^{n}\sum_{j\neq i}^{n}\phi(z_{T_i},z_{T_j}), \quad (4)\end{aligned}$$

where, m and n are the number of samples drawn from Z_S and Z_T, and ϕ is a Gaussian kernel representing the feature mapping transformation. A minimisation of MMD ensures that the distributions Z_S and Z_T are similar across each statistical moment, which aids in making predictions about the unlabelled Target domain.

2.1.3. Domain adaptation-oriented loss function

The Negative Log Likelihood, NLL, cost is used to penalise incorrect classification predictions made by the model and is given by

$$\text{NLL}(\theta) = -\sum_{i=1}^{k}\left(y_i\ln(\hat{y}_{\theta i}) + (1-y_i)\ln(1-\hat{y}_{\theta i})\right), \quad (5)$$

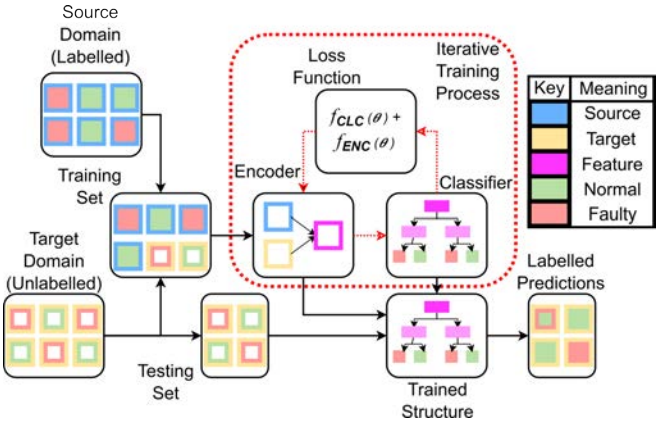


Figure 1. An example of an unsupervised domain adaptation approach. An encoder and classifier are trained simultaneously to generate both a representative feature space and accurate class predictions.

where, θ is a set of probabilities attributed to each class prediction, k is the number of predictions made, y is the true class of each sample and $\hat{y}$ is the predicted class.

To perform domain adaptation the following loss function is used:

$$\mathcal{L}_{DA} = \text{NLL}(\theta) + \text{MMD}(Z_S,Z_T) - \text{MI}(Z_T,Z). \quad (6)$$

This is a common form for a loss function seen in an unsupervised domain adaptation setting, visualised in Figure 1.

For comparison, this work will also use a simplified version of this loss function, $\mathcal{L}_S = \text{NLL}(\theta)$, to represent a loss function used by a conventional supervised learning approach.

3. Methodology

3.1. Data generation

In light of the lack of labelled relevant NPP accident data, RELAP5, a nuclear reactor modelling and simulation tool, was used to generate the data used in this work. Rather than perform domain adaptation between model generated data and real data collected from a plant, a high-fidelity physics models is used under three different configurations that represent data domains with varying levels of divergence between them. The model configurations represent 1) A large four-loop 3600MW civil nuclear plant with nominal historical usage, 2) A similar large plant with a greater historical power usage and 3) A small two-loop 50MW with nominal historical usage.

3.1.1. Model modifications

To replicate the full range of operating conditions of a civil nuclear plant, and for a machine learning framework in this

setting to be trained robustly, data containing examples of dynamic events are provided. These events come not only from the presence of LOCA/faults, but also reflect dynamism in the normal operation of a plant, such as reactivity insertion.

The specification of a general table used to define the core reactivity or power (depending upon the reactor kinetics model used by the script) proved sufficient for providing the kind of input-derived transient events required. Typically, these input demands have magnitude between 1-10% of the input reactivity/steam off-take of that observed at the reactor's steady state rated power output level.

A small degree of Gaussian noise was added to the input reactivity demand profile to simulate process noise. The scale of the input noise was less than 10% of the magnitude of the changes in input demands. Set-points and thresholds that define variable and logical trips for control systems in the model were perturbed to emulate differing operator characteristics on each run.

Each simulation was performed with or without the presence of a LOCA (hence being classed as 'Normal' or 'Faulty').

Breaks were inserted into the primary circuit of the reactor coolant system to simulate a LOCA. Breaks are simulated at the inlet and outlet of the hot and cold legs of the primary circuit, as well as at the outlet of the steam generator in the secondary circuit. All breaks used the counter-current flow model, with standard choking flow. The full abrupt change model was used meaning that all breaks occurred instantaneously, rather than develop throughout the course of one sample of time-series data. The breaks are modelled as a valve with given cross-sectional area. The cross-sectional area of the break-valve is adjusted to define the size of the break relative to the cross-sectional area of the pipe to which the break-valve is located. The break sizes are uniformly sampled in the range $[0.02\%, 0.2\%]$ of the area of the pipe for the 3600MW plant, and the range $[0.1\%, 1\%]$ of the area of the pipe for the small 50MW plant, representing very small breaks. Each simulation is run for 1000 seconds.

A summary of the numerical changes to the high-fidelity physics models is as follows:

- Transient operating power provided by control of reactor rod position or steam off-take at the steam generator. Operating power level varied between +-10% of the rated capacity of each plant.
- Gaussian process noise inserted with transient input signals, scaled to +- 1% of the rated capacity of each plant.
- Control system thresholds shifted by -2%, +2% or unchanged for each simulation.
- Breaks inserted with magnitudes in the range $[0.02\%, 0.2\%]$ and $[0.1\%, 1\%]$ of the cross-sectional area of the pipe in which they are located for the 3600MW and 50MW plant respectively.
- Each simulation is run for 1000 seconds.

3.1.2. Differences between models (domain divergence)

This work focuses on exploring the implication of an increasing divergence between data domains. In this context, this requires multiple high fidelity physics models from which to gather data. The first of the template models used describes a large Four-loop 3600MW PWR with characteristics designed to be a 'fictitious approximation' of values present in a Westinghouse plant.

To provide an example of a relatively small domain divergence, the 3600MW PWR model is used to provide data representative of the same plant at different stages in its operating cycle. To achieve this, different model initialisation applied that define different average operating power output for the first year of operation. An 'Underworked' version of this Four-loop plant was defined to have operated at 2400MW (significantly less than the 3600MW rated capacity) for its first year of operation. This model was used as the 'Source' domain model. A 'Typically worked' version of the same plant is defined to have operated at its rated capacity of 3600MW for the first year of it's operation. Additionally, pump speeds throughout the primary and secondary circuits are increased, allowing for different dynamics to manifest in this version of the plant. The model generated data used for the 'Target 1' domain.

A second, smaller Two-loop 50MW PWR plant was chosen to represent a more drastic domain change. This plant is simulated with the same relative changes in input reactivity and trip set-point attitudes, but possessing disparate dynamics and steady state behaviour owing to the different physical properties of the smaller plant. Data generated from this model is used for the 'Target 2' domain.

A quick understanding of the degrees of divergence between these domains can be gained by observing the first principal component of a principal component analysis performed on each domain, shown in Figure 2. The distribution of the first principal component differs greatly depending upon which domain the data comes from, although the difference between the Source and Target 2 domains is much larger than that between the Source and Target 1 domains. This domain divergence would cause a conventional supervised learning approach to suffer, due to the difficulty in transferring knowledge between disparate domains. Additionally, note the large degree of overlap that exists between 'Normal' and 'Faulty' classes of data in each domain. This implies that, since principal components analysis is a linear technique, a linear classifier would struggle to separate samples of data from different classes.

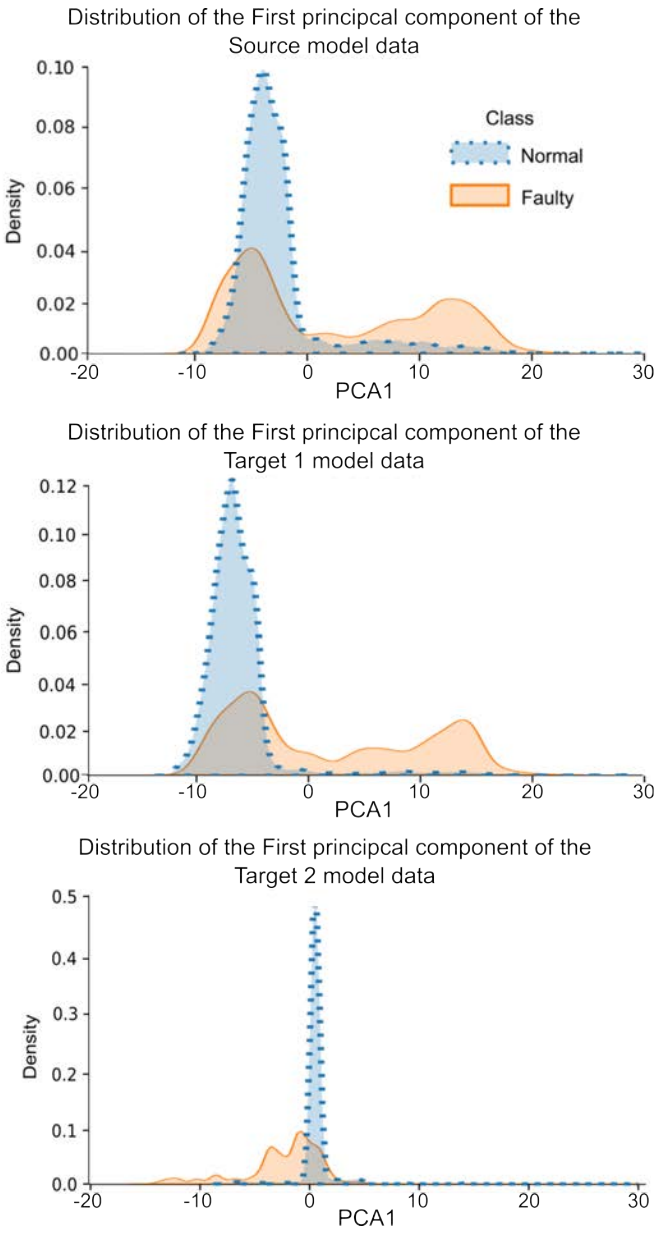


Figure 2. The distributions of the first principal components of the data from each of the three domains. Note the larger difference in distributions between the Source and Target 2 domains.

3.1.3. Processed data structure

Observations of the plants were made by simulating sensors in locations throughout both the primary and secondary circuits, listed in Table 1. In total, 15 data streams were extracted from each simulation, representing pressures, temperatures, mass-flow rates and valve states, with locations shown in Figure 3.

In addition to these measured values, the input reactivity profile (the power demand) and the reactor output power were supplied as part of the data vector provided to the domain adaptation network. Each batch of data the encoder receives is made of samples derived from both the Source and Target domains. At the encoder, these samples are unlabelled. Each sample x is an $N \times T$ vector representing T time-steps of N sensor readings. The input dimension of the encoder is $B \times N \times T$, where B is the number of samples used per batch.

Table 1. A list of the measured values used in this work.

Value specification		
ID	***Description***	***Units***
0	Cold-leg Coolant Pressure	P
1	Cold-leg Inlet Coolant Temperature	°C
2	Cold-leg Outlet Coolant Temperature	°C
3	Hot-leg Coolant Pressure	P
4	Hot-leg Inlet Coolant Temperature	°C
5	Hot-leg Outlet Coolant Temperature	°C
6	Pressuriser Relief Valve State	-
7	Main Steam Isolation Valve State	-
8	SG Feedwater Regulating Valve State	-
9	Cold-leg Coolant Mass Flow-rate	kg/s
10	Hot-leg Coolant Mass Flow-rate	kg/s
11	Reactor Coolant Pump Mass Flow-rate	kg/s
12	SG Feedwater Inlet Mass Flow-rate	kg/s
13	Pressurizer Inlet Mass Flow-rate	kg/s
14	Charging Pump Mass Flow-rate	kg/s

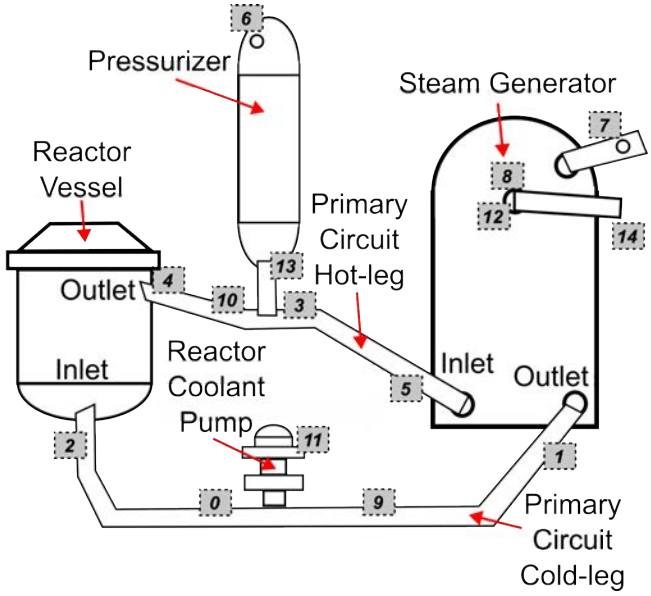


Figure 3. A simplified view of the primary coolant circuit of a PWR. The measured sensor values used in this work have their sensed locations numbered.

3.2. Model architecture

The network described in this work consists of an encoder, tasked with extracting generalised feature maps from the raw input vector data, and a classifier aimed at detecting the presence of 'Faulty' data. Different model architectures were were trialed and they are described in detail below.

3.2.1. Encoder variations

To investigate the impact that the internal structure of the feature-generation stage of the network has in the context of NPP time-series data, several forms of encoder were considered, with increasing degrees of 'connectivity' between different input sensor readings, visualised in Figure 4.

Baseline Encoder: Separate 1D convolutions

Two kernels per input sensor channel, with kernel grouping number set equal to the input dimension at each point in the encoder. This has the effect of training kernels without combining information from multiple sensor channels simultaneously. Kernels act along the time dimension of each sensor reading.

Aggregate Encoder: Summed 1D convolutions

Grouping number for convolutional layers set to 1, meaning kernels are passed along a single time-series sensor reading, before being combined in a weighted sum to generate a convolution which contains information from each sensor measurement simultaneously.

Recurrent, Fully-Connected Encoder: Gated Dense 1D convolutions

The summed 1D convolutions have been performed as above followed by a fully-connected layers Additionally, a gated recurrent unit (GRU) layer is used before the second set of convolutions.

These modifications attempt to allow the encoder to efficiently consider long-term dynamics that may be important to fault detection in this context.

3.2.2. Classifier & loss function

The classifier is shared by each of the different encoder variations described above. The classifier consisted of a series of fully connected layers followed by batch normalisation layers, shown in Figure 5. A dropout layer is included to encourage regularisation and avoid over-fitting.

4. Experimental validation

The loss function used by each model varies depending upon whether domain adaptation is used. When a model is applied without domain adaptation, the loss $\mathcal{L}_S$ is used. The loss $\mathcal{L}_{DA}$ is used when domain adaptation is required. The 'Baseline' model was tested in each data domain twice, once using $\mathcal{L}_S$ and once using $\mathcal{L}_{DA}$. The Aggregate Encoder and Recurrent Fully-Connected Encoder were tested on all data domains using the loss $\mathcal{L}_{DA}$.

4.1. Hyper-parameter tuning result

The model hyper-parameters were tuned heuristically for each model variant. Hyper-parameters were chosen as

Baseline Model:

- Learning rate: $8e-4$

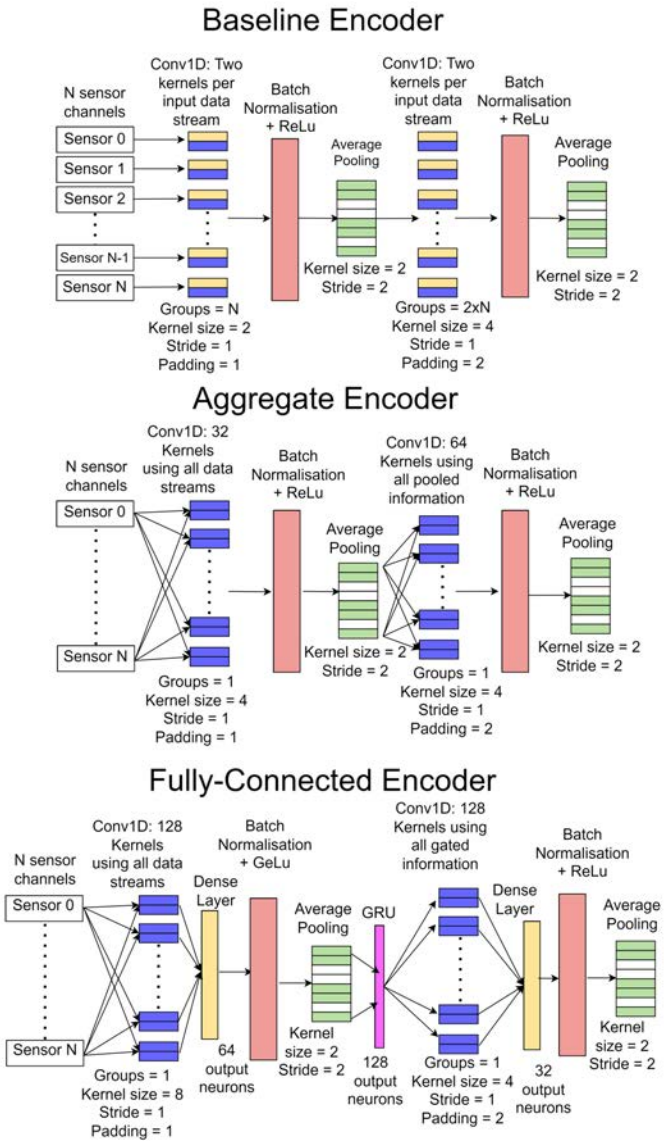


Figure 4. Architectures of three different encoders used in this work, with increasing degrees of connectivity between the input sensor readings.

- Dropout percentage for classifier input: $7.0\ \%$
- Scaling factor for MMD term in loss function: 0.5

Aggregate Encoder Model:

- Learning rate: $9e-4$
- Dropout percentage for classifier input: $4.1\ \%$
- Scaling factor for MMD term in loss function: 0.5

Fully-Connected Encoder Model:

- Learning rate: $6e-4$
- Dropout percentage for classifier input: $5.6\ \%$
- Scaling factor for MMD term in loss function: 0.5

4.2. LOCA detection

This section, detailing the results from this work, is divided into three parts covering, binary LOCA classification performance, detection performance by break size and detection performance by break location.

4.2.1. Binary classification performance

When tested on the the Source domain, each model performed similarly in classification accuracy, with over 93% of the samples observed being correctly classified as either 'Normal' or 'Faulty' for each model variant, as shown in Table 2. Without the necessity for domain adaptation in this case, each model was able to create a robust feature map of the distribution of data observed in the Source domain. The increased complexity of the Aggregate and Fully-Connected Encoders offered little to no benefit in this conventional supervised setting, with the training and testing sets being drawn from the same domain.

Disparities in model performance start to appear when the testing set is drawn from the Target 1 data domain. This testing set represents a slight shift in data distribution from the Source domain training set, which reveals the importance of the inclusion of model architectures specifically designed to aid in domain adaptation. The models utilising the more sophisticated domain adaptation-oriented loss function retain the majority of their classification performance when compared to test results from the Source domain, whilst the Baseline supervised model, using $\mathcal{L}_S$, suffers a sizeable reduction in classification accuracy.

This difference in performance is owed to the fact that the domain adaptation-oriented loss function contains terms that inform the encoder about the statistical properties of the generalised feature space that it is tasked with creating. Consideration of the MI between the entire feature space and the encoded representation of the Target 1 domain data helps to reduce the likelihood of encountering unlabelled Target 1 domain samples during testing that do not possess some information that the model has previously observed during training. Prompting this overlap in shared information increases that chances that the model will hold some 'relevant' feature space representation for these 'unique' unseen phenomena, which is crucial due to the high variability of the generated data. Additionally, minimisation of the MMD at the encoder aids the classifier with inferring class labels belonging to the unlabelled Target 1 domain data. This is explained by the fact that a reduction in MMD between the encoded representations of the Source and Target 1 domains implies that samples belonging to one class (for example, 'Faulty') that exist in one region within the encoded Source domain feature space, should exist in a similar 'relative' location within the encoded Target domain feature space. It is through this knowledge transfer that Source domain information can be leveraged to support Target domain class predictions.

The final domain shift, between training on Source domain model data and testing on the Target 2 domain data represents a more severe divergence between domains. As is evident from the average test accuracies, the conventional supervised model fails to bridge this gap between differing data distributions and records a poor performance of less than 50% classification accuracy. It is at this magnitude of domain divergence that the increased complexity of the Aggregate and Fully-

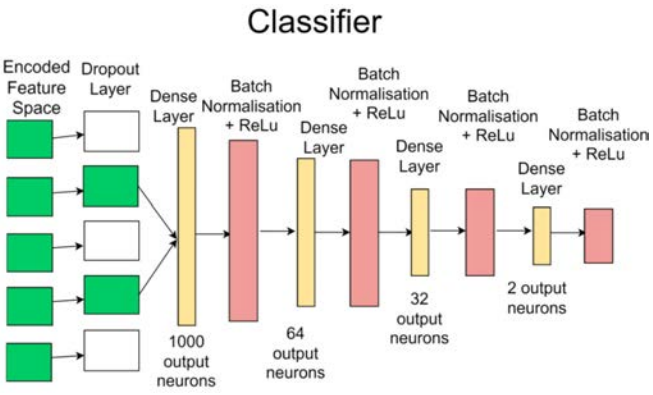


Figure 5. Architecture of the classifier used in this work.

Table 2. Average test accuracies (%) of each model type on each domain set, after being trained on the Source data domain.

Model Type	Source Domain	Target 1 Domain	Target 2 Domain
Baseline Supervised (No DA)	93.41	78.02	45.57
Baseline With DA	-	91.24	81.46
Aggregate Encoder	93.96	91.02	89.85
Fully-Connected Encoder	93.22	91.27	89.57

Connected Encoders demonstrate value. Although the DA-focused loss function is enough to restore the majority of the lost performance for the baseline model, the less-connected Encoder lacks the ability to consider the more disparate dynamics evident in the Target 2 domain.

In the context of this work, allowing the encoder to combine input sensor channels at the first convolutional layer before performing the convolution (such as in the Aggregate and Fully-Connected encoders), may allow the models to exploit class-specific relationships between sensed quantities. For example, in the presence of a LOCA, a brief divergence of primary circuit hot-leg temperature and pressure may occur. If these sensed values were provided to the same convolutional kernel at the input layer of the encoder, then the kernel could utilize this relative disparity to generate a recognisable identifier of a LOCA class sample. This improvement in performance compared to the 'Baseline With DA' model suggests that there is more information available with respect to the problem of LOCA detection if the sensed values are processed relative to each other, rather than processed in parallel.

Although the nature of LOCA simulated in this work vary drastically in magnitude and location within the modelled plants, there exists the possibility for other fault cases to transpire in an NPP. Without explicit knowledge of the existence of these faults (other than LOCA), the accurate classification of these samples as 'Faulty' would depend upon the similarity of these encoded samples to the 'Normal' encoded data. The performance of the models in this work on LOCA-specific fault detection is good, meaning the classifier used can differentiate between 'Normal' data, and all other LOCA data. If another fault case, previously unseen by the models described, manifested in a similar fashion to a LOCA, it is likely it would be identified as 'Faulty'. However, as can be observed in later analysis on model performance by break location, there can be large difference in classification accuracy for a single model across faults from multiple locations, so it would not be reliable to depend upon these methods as part of generic 'anomaly detection' techniques.

An understanding of the impact of the encoder in this work can be understood if the distributions of the encoded data from each domain are observed, shown in Figure 6. The data used in this figure are drawn from the Fully-Connected encoder. Viewing the first principal component of the post-encoder data from each domain reveals that the three domains appear much more similar to each other once expressed in the generalised feature space the encoder provides. As observed previously, the encoded distributions still share a large degree of overlap between 'Normal' and 'Faulty' data. Since principal components analysis is a linear transformation, this suggests that a linear classifier would struggle to reliably predict the class of unlabelled samples, and that the models used in this work which perform well must consider nonlinearities in the data.

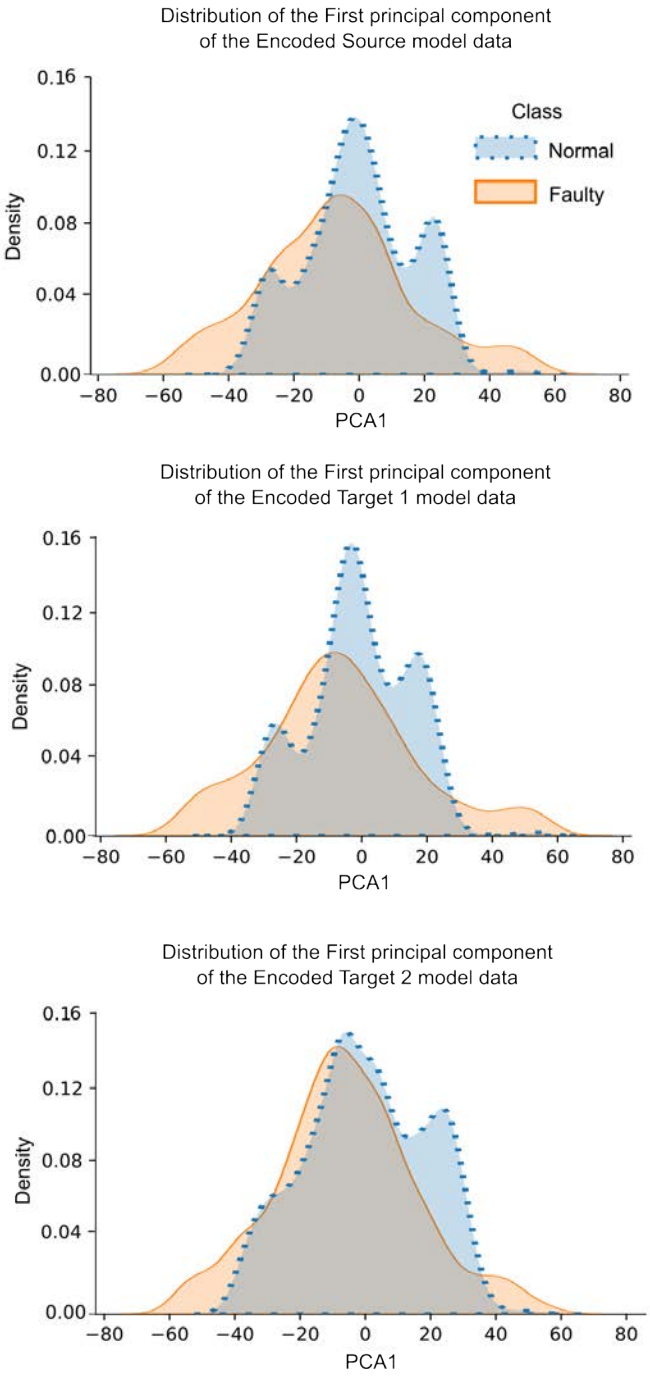


Figure 6. The distributions of the first principal components of the encoded data from each of the three domains. Note the more similar distributions compared to the first principal component of the input space.

4.2.2. LOCA detection by break size

An interesting perspective by which to consider the performance of these models in the context of NPP fault detec-

tion is to observe only the primary circuit breaks and identify the thresholds above which each model can always identify a LOCA. LOCA detection by break size results are shown in Figure 7.

In the Source domain setting, each model could reliably categorise fault data with break size above 0.1% of the pipe cross-sectional area as 'Faulty'.

The performance is retained when the models are tested on the Target 1 domain set, however the rate at which the baseline model without domain adaptation can successfully categorise samples below 0.1% is substantially lower than in the source domain.

The performance loss is exaggerated as the gap between domains increases further still: without considering domain adaptation, there is no size of primary circuit break that the baseline model can always categorize correctly as 'Faulty'. With the only alteration being the inclusion of domain-adaptation focused terms in the loss function, the Baseline model (With DA) can, on average, identify 20% more faults successfully.

The other models retain a substantial proportion of their ability to successfully categorize all break sizes, even in this most extreme domain divergence example.

4.2.3. LOCA detection by break location

The 'Faulty' samples in this work were not only drawn from a range of possible sizes, but also placed at varied locations throughout the primary and secondary circuit of each PWR. Primary circuit breaks occur at either the inlet or outlet of the hot or cold legs. Secondary circuit breaks were inserted at the outlet of the steam generator for the respective loop. LOCA detection by break location results are shown in Figure 8.

When tested in the native Source domain setting, both the Baseline Supervised and Aggregate Encoder models correctly classify all nominal operation data, along with a consistently high successful classification rate of primary circuit breaks as 'Faulty'. The Fully-Connected Encoder sacrifices the successful classification of a small number of 'Normal' samples in order to correctly identify each primary circuit break observed in this testing environment as 'Faulty'.

Classification of breaks located at the steam generation outlet was comparatively poor. This is perhaps due to the lower number of observed examples of secondary circuit breaks during training, or the potential for secondary circuit breaks to be harder to identify under the lower-fidelity of the secondary circuit physical model in comparison to the primary circuit. Only the Fully-Connected Encoder model is able to correctly classify any of these samples as 'Faulty', which suggests that secondary circuit breaks appear more similar to 'Normal' operational data from the perspective of these classifiers.

In the Target 1 domain, representing a slight shift in domain

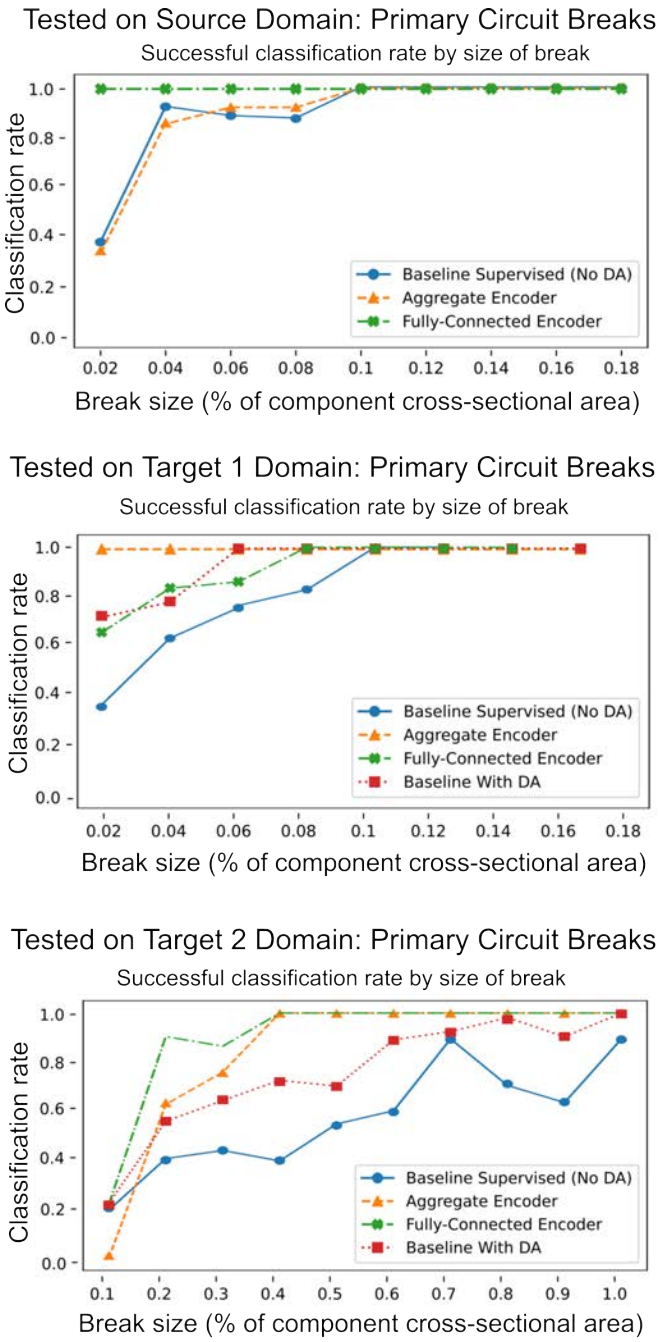


Figure 7. Successful classification rate by fault break size for each model combination, when tested on each domain.

distribution from the source domain, the Baseline Supervised model retained its ability to identify 'Normal' operational data, whilst 'Faulty' sample classification accuracy degraded. As in the Source domain, the Baseline Supervised and Aggregate Encoder models were not able to correctly identify any secondary circuit breaks as 'Faulty'. The Baseline With DA model (using $\mathcal{L}_{DA}$) gives some improvement in primary cir-

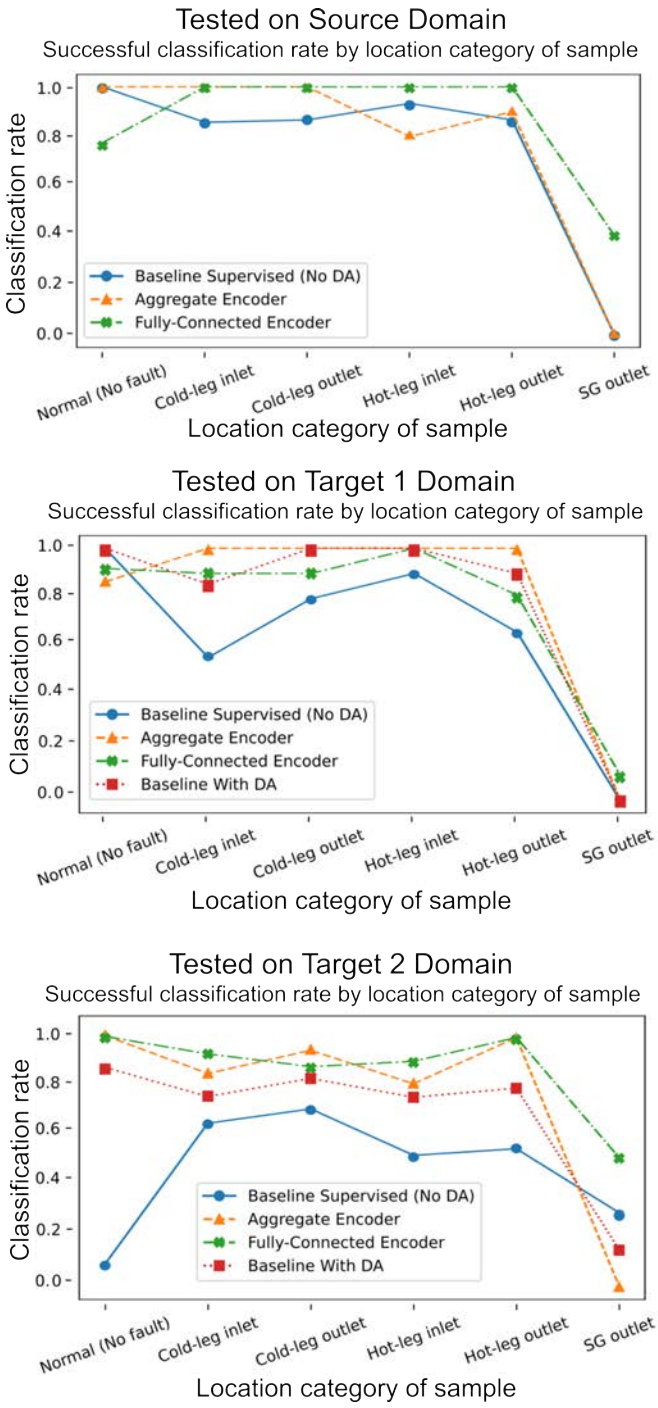


Figure 8. Successful classification rate by fault location for each model combination, when tested on each domain.

cuit break identification, but remains unable to recognise the 'Faulty' class of secondary circuit break samples. As in the previous testing domain, the Fully-Connected Encoder is the only model capable of correctly classifying any steam generator outlet samples, albeit a very small proportion of the samples it observed.

In the most extreme example of domain divergence between the Source domain and the Target 2 domain, the classification ability of the Baseline Supervised model, using $\mathcal{L}_S$, deteriorates. In this unfamiliar domain, the Supervised model is barely able to correctly classify any 'Normal' data. The Baseline model using $\mathcal{L}_{DA}$ restores some classification ability of 'Normal' samples, and improves the detection of 'Faulty' samples in all locations except the secondary circuit. Once again, the Fully-Connected Encoder displays the best detection performance for secondary circuit breaks. This indicates the importance of combining the sensor channels at the Encoder level in order to generate generalised features which can remain relevant between disparate training and testing domains.

5. Conclusion

The results detailed in this work highlight the value in incorporating domain adaptation techniques in scenarios where discrepancies exist between the training and testing data domains. Even when the scale of these discrepancies can become large, fundamental DA concepts provide a significant improvement in performance when compared to a conventional supervised learning approach. Additionally, the results draw attention to the importance of combining input sensor channels in this context. Models which consider information from multiple sensed sources simultaneously during their construction of each encoded feature map retained a much greater proportion of their classification ability seen in the Source domain classificaiton performance.

Acknowledgment

This work was partly funded by the Aerospace Technology Institute under the REINSTATE project.

References

Aldemir, T. (2013). A survey of dynamic methodologies for probabilistic safety assessment of nuclear power plants. *Annals of Nuclear Energy*, *52*, 113-124. (Nuclear Reactor Safety Simulation and Uncertainty Analysis)

Chen, J., Wang, J., Zhu, J., Lee, T. H., & de Silva, C. W. (2021). Unsupervised cross-domain fault diagnosis using feature representation alignment networks for rotating machinery. *IEEE/ASME Transactions on Mechatronics*, *26*(5), 2770-2781.

Farber, J. A., & Cole, D. G. (2020). Detecting loss-of-coolant accidents without accident-specific data. *Progress in Nuclear Energy*, *128*, 103469.

Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W.-K., & Yang, H. (2020). Status of research and development of learning-based approaches in nuclear science and engineering: A review. *Nuclear Engineering and Design*, *359*, 110479.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., &

Muller, P.-A. (2019, March). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963.

Lee, G., Lee, S. J., & Lee, C. (2021). A convolutional neural network model for abnormality diagnosis in a nuclear power plant. *Applied Soft Computing*, *99*, 106874.

Mahmoodi, R., Shahriari, M., Zolfaghari, A., & Minuchehr, A. (2011). An advanced method for determination of loss of coolant accident in nuclear power plants. *Nuclear Engineering and Design*, *241*(6), 2013-2019. ((W3MDM) University of Leeds International Symposium: What Where When? Multi-dimensional Advances for Industrial Process Monitoring)

Qian, Q., Qin, Y., Luo, J., Wang, Y., & Wu, F. (2023). Deep discriminative transfer learning network for cross-machine fault diagnosis. *Mechanical Systems and Signal Processing*, *186*, 109884.

Rivas, A., Delipei, G. K., Davis, I., Bhongale, S., & Hou, J. (2024). A system diagnostic and prognostic framework based on deep learning for advanced reactors. *Progress in Nuclear Energy*, *170*, 105114.

Sakurahara, T., O'Shea, N., Cheng, W.-C., Zhang, S., Reihani, S., Kee, E., & Mohaghegh, Z. (2019). Integrating renewal process modeling with probabilistic physics-of-failure: Application to loss of coolant accident (loca) frequency estimations in nuclear power plants. *Reliability Engineering & System Safety*, *190*, 106479.

Wang, H., jun Peng, M., Ayodeji, A., Xia, H., kun Wang, X., & kang Li, Z. (2021). Advanced fault diagnosis method for nuclear power plant based on convolutional gated recurrent network and enhanced particle swarm optimization. *Annals of Nuclear Energy*, *151*, 107934.

Wei, L., & Keogh, E. (2006). Semi-supervised time series classification. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (p. 748–753). New York, NY, USA: Association for Computing Machinery.

Zhang, J., Li, X., Tian, J., Jiang, Y., Luo, H., & Yin, S. (2023). A variational local weighted deep sub-domain adaptation network for remaining useful life prediction facing cross-domain condition. *Reliability Engineering & System Safety*, *231*, 108986.

Zhang, Y., Ren, Z., Feng, K., Yu, K., Beer, M., & Liu, Z. (2023). Universal source-free domain adaptation method for cross-domain fault diagnosis of machines. *Mechanical Systems and Signal Processing*, *191*, 110159.

Domain Adaptation *via* Simulation Parameter and Data Perturbation for Predictive Maintenance*

Kiavash Fathi[1,2,**], Fabio Corradini[3,**], Marcin Sadurski[1,***], Marco Silvestri[3,***],
Marko Ristin[1], Afrooz Laghaei[1],
Davide Valtorta[4], Tobias Kleinert[2], Hans Wernher van de Venn[1]

[1] *Institute of Mechatronic Systems, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland*
fath@zhaw.ch, sadu@zhaw.ch, rist@zhaw.ch, lagf@zhaw.ch, vhns@zhaw.ch

[2] *Chair of Information and Automation Systems for Process and Material Technology, RWTH Aachen University, 52064 Aachen, Germany*
kiavash.fathi@rwth-aachen.de, kleinert@plt.rwth-aachen.de

[3] *Department of Innovative Technologies, University of Applied Sciences of Southern Switzerland, SUPSI, Via la Santa 1, 6962, Lugano-Viganello, Switzerland*
fabio.corradini@supsi.ch, marco.silvestri@supsi.ch

[4] *SAECON Sagl - Safety and Engineering Consulting, via Grancia 1, 6926 Montagnola, Switzerland*
valtorta@saecon.net,

ABSTRACT

Conventional data-driven predictive maintenance (PdM) solutions learn from samples of run-to-failures (R2F) to estimate the remaining useful life of an asset. In practice, such samples are scarce or completely missing. Simulation models can be oftentimes used to generate R2F samples as a replacement. However, due to the complexity of the assets, creating realistic simulation models is tedious, or even impossible. Thus generated R2F data cannot be used to create reliable PdM models as they are highly sensitive to noises in the sensors or small deviations in system working condition. To address this, we present a new concept of simulation data generation based on supervised domain adaptation for a regression problem where the remaining useful life (RUL) or the health index (HI) of the system is predicted. Apart from input and output domain shift, given the changes in the dominant failing component and its degradation process, the function mapping sensor readings to RUL and/or HI is also prone to changes and thus is a random process itself. Therefore, we aim to generate R2F training data from different working conditions and possible failure types using parameter randomization in the simulation model. By sampling from various configurations within simulation model's parameter space, we ensure that the trained data-driven PdM model's performance is not impacted by the initial conditions and/or the changes in the degradation of the system's condition indicators. Our results indicate that the model is robust to signal reading manipulation and showcases a more spread-out feature importance across a wider range of sensor readings for making predictions. We also demonstrate its applicability on the real-world factory physical system whilst our models were mainly trained using generated data.

1. INTRODUCTION

Accurate prediction of a production asset's health state enables effective implementation of a predictive maintenance (PdM) solution. Such a solution can help reduce both the cost and occurrence of unscheduled maintenance of the targeted production asset (Cui, Du, & Hawkes, 2012; Rahat et al., 2022). With the advancements in sensor technologies, data acquisition and analysis, numerous PdM solutions predict the remaining useful life (RUL) and/or the health index (HI) by either date-driven models, model-based models or hybrid of the two.

* This research was supported by Innosuisse - Swiss Innovation Agency, Innosuisse Grant no: **58020.1 IP-ENG**, entitled **MaintAIn - AI supported hybrid Predictive Maintenance**.
** These authors contributed equally.
*** These authors contributed equally.

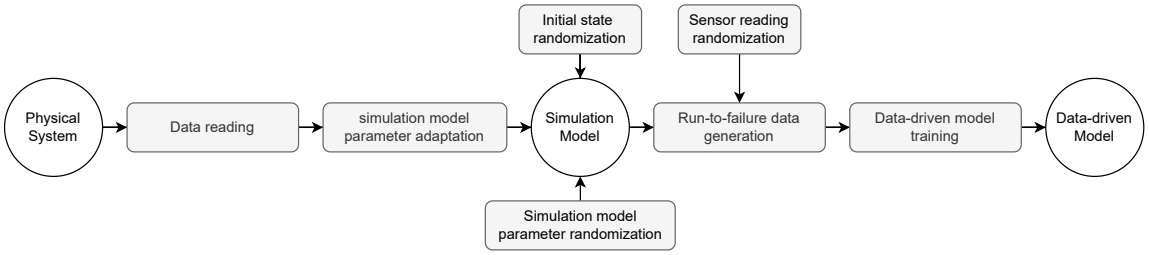


Figure 1. Overview of the proposed method

The *data-driven models* use historical data to train the model of RUL, so their quality depends on the data quality. It is thus of utmost importance that the training data meets the minimum data quality requirements (Liu, Wang, Ma, Yang, & Yang, 2012). However, it is rarely the case that the available data from a production asset not only has enough samples of failure, but also covers all possible failure types of the system (Fathi, van de Venn, & Honegger, 2021).

This in turn raises the need for the data generation via simulation models, which underlines the importance of currently missing related work on simulation-to-real transfer and domain adaptation (DA) techniques for RUL and/or HI estimation in PdM.

The *model-based approaches* leverage mathematical and physical models to estimate the RUL. This usually requires parameter tuning, *e.g.,* using Markov process model or Winner process, for converging to the behavior of the physical system (Hanachi, Liu, Banerjee, Chen, & Koul, 2014; Si, Wang, Hu, Zhou, & Pecht, 2012; Thelen et al., 2022). The parameter tuning is highly sensitive to the parameter initialization, which is normally based on the empirical knowledge from the physical system (Lei et al., 2016). Even when the system parameters are estimated correctly, any changes in the production setting and/or the production asset itself requires a re-calibration.

Both the data-driven and the model-based approaches suffer from inadequacy in practical solutions: the relevant data is either missing, or the models are not robust enough, respectively.

Hybrid PdM models combine the two approaches by learning from both the historical data and the data synthetised from the simulation models. The hybrid models have proven to be effective in terms of reliability and efficiency, and address some of the issues of pure data-driven or model-based solutions, such as reduction in data acquisition time and increased model robustness (Chang, Fang, & Zhang, 2017; D. Chen et al., 2022; Lin, Yu, Wang, Che, & Ni, 2022; Didona & Romano, 2014). In practice, the sampling from simulation models can not come up for the lack of historical data, so the conventional hybrid PdM models are also heavily dependant on annotated data. They are normally applied to systems for which extensive labelled datasets are available, *e.g.,* RUL prediction for lithium-ion batteries and aircraft engine (Fei, 2022; Saxena, 2023). As large datasets are not available for production assets in a manufacturing factory, the conventional hybrid models are inapplicable in the manufacturing setting.

In this paper, we focus on hybrid modelling for this specific scenarios where gathering extensive labelled data from the physical system is non-trivial (or even impossible), and where precise and robust simulation models are unavailable. To compensate for that, we propose to extract as much value as we can from both sources by applying the *domain adaptation* between the simulation and the real-world scenarios (see Fig. 1).

In the same vein, we further improve the robustness by considering the data distribution shifts which are a common consequence of diverse manufacturing requirements in an Industry 4.0 setting due to flexible and adaptable production (Fathi, Sadurski, Kleinert, & van de Venn, 2023). In order to cover as much of the parameter space of the system as possible, we alter accordingly the modelling of the initial condition and the degradation of the system.

The main contribution of this paper is thus four-fold. We:

1. Propose a hybrid PdM solution which relies mainly on the data from a simulation sub-module and few samples from the target domain for training its data-driven sub-module (supervised DA),
2. Propose a new concept of simulation data generation aiming for domain adaptation called Parameter and Data Perturbation (PDP), for covering as much of the parameter and degradation space of the physical system as possible,
3. Inspect the impact of the changes in the generated data from the simulation sub-module and the physical system on the performance of the data-driven sub-module,
4. Demonstrate how the additional simulation data used for model training results in a more spread-out feature importance across a wider range of sensor readings from the system while making predictions.

The rest of the paper is outlined as follows. First, some related work addressing synthetic data generation from simula-

tion models for dealing with scarce labelled data and domain adaptation are presented. Afterwards, the details of the simulation model are provided. Thereafter, the results of model training using the simulation model data with PDP are presented. Lastly, discussion and the future work of this work are presented and some conclusions are drawn.

2. Related Work

2.1. Lack of annotated data from the target system

One method used for reducing the time spent gathering data from the target application using hybrid modelling is *boostrapping* (Didona & Romano, 2014). The main idea of *boostrapping* is to rely on a simulation model of the target use case and to generate initial synthetic training set for the data-driven model training. Thereafter, the data-driven model tries to incorporate knowledge from the target system as soon as a data point is available. In (Didona & Romano, 2014), the authors propose to remove the synthetic data points in the vicinity of samples from the target system to prevent obfuscating information from the real samples. However, for the purpose of RUL prediction in PdM, the annotated samples from the physical system are scarce and costly to gather. Hence, we propose to instead to keep these valuable samples and to combine them with data from different working conditions of the system generated from the simulation model

2.2. Adaptation to different working condition

Another important issue impacting the performance of data-driven PdM models is the varying working condition of the production assets in industry. These constant changes can make models trained with a specific working condition (a.k.a. source distribution) obsolete as changes occur in the system (Ragab, Chen, Wu, Kwoh, & Li, 2020). They cause a data distribution shift between the data employed to train the PdM model and the data acquired during the model's deployment in the production line. This discrepancy between the source and target distribution raises the need for techniques such as domain adaptation. In fact, domain adaptation aims to train a model on multiple source domains which are annotated so that the model can generalized to new and unseen target domains (Farahani, Voghoei, Rasheed, & Arabnia, 2021).

To the best of our knowledge, no other PdM work adopted domain adaptation methods for robust RUL and HI estimation in light of lacking annotated historical data from the physical system using simulation-to-real transfer techniques. Moreover, the current literature (Farahani et al., 2021; Yu, Fu, Ma, Lin, & Li, 2021; Rahat et al., 2022; Yang, Lei, Jia, & Xing, 2019; Gao, Liu, Huang, & Xiang, 2021; Wang, Taal, & Fink, 2021) reduce PdM to be a binary or a multi-class variable. We treat it as estimating the RUL or HI as a continuous value for better adaptability to different scenarios.

In addition, given the degradation process of different critical components of a production asset, the labelling function mapping the input space to the output space, can not only be different in the target domain, but also change in the target domain given the dominant degradation process of any arbitrary critical component (Cortes & Mohri, 2011; Nejjar, Geissmann, Zhao, Taal, & Fink, 2024)

These two, the estimation of RUL and/or HI as a continuous variable under the assumption of scarcity of annotated data from the target domain, and the possibility of data distribution shifts in the target domain are the primary motivation behind the proposed PDP method outlined in this work (Fig. 2).

2.3. Simulation-to-real transfer

Numerous application, especially safety-critical system, suffer from lack of labelled data as gathering such datasets is costly or endangers the human operator (Kaufmann et al., 2020; Tiboni, Arndt, & Kyrki, 2023). Therefore, simulation models are used to recreate different scenarios which are also labelled for model training. Nonetheless, modelling errors and the complexity of physical systems prevent the zero-shot deployment of data-driven models trained with such simulation models. One way for increasing the robustness of the trained models is randomize the dynamic parameters of the system which is a.k.a. *domain randomization* (Peng, Andrychowicz, Zaremba, & Abbeel, 2018). Doing so increases the robustness of the trained model at the the cost of its optimality (Tiboni et al., 2023). In this paper, we use the same method to cover as much as the parameter and degradation dynamics space as possible. Different starting conditions, model parameter and degradation processes are the main sources of domain randomization in this paper (Fig. 2).

3. Simulation model

Data-driven monitoring systems require continuous data collection that must extend over a period of time before they can provide effective results (Bonomi et al., 2021). Such data collections, referred to as run-to-failure (R2F) data, are normally expected to start from a healthy production asset state and end with the asset failing or malfunctioning. In the present scenario, there is the additional consideration that R2F tests are, by their nature, long-lasting while accelerated destructive testing is not always possible. Additionally, these tests are expensive with uncertain success. We propose to use a simulated model of the system to obtain realistic data (Ferrario et al., 2019) on its response to the system's most common types of wear and tear.

The model creation phase is critical because it must satisfy several conflicting requirements:

- The model must be complex enough to represent deteriorating operating conditions realistically.

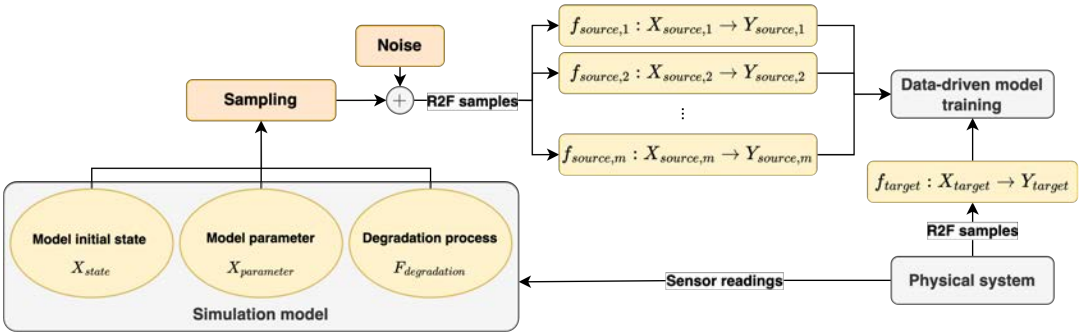


Figure 2. DA via PDP

- The model must be simple enough to be calibrated in a short time and from accessible data.
- The model must take into account the system whose condition is being monitored, the measurement instruments used, and how the data are processed. To some approximation, it must also take into account the context in which the system operates (*e.g.*, the effect of other components).

To meet all these needs, we developed a system with the following characteristics:

- Our models follow the multi-physical system (*Simscape*) as a concentrated variable model. This allows parameters to be easily configured.
- The model is wrapped in a Python script that allows generating the random parameters, handles post-processing, and eventually repeats or recovers the simulation scenarios in case of failure.

3.1. Overview of the simulation model

Our specific use case is the condition and life monitoring of a series of pneumatic cylinders. Based on its characteristics, we separate the model into 5 macro blocks (Fig. 3):

1. **Control system block:** modeled as timed output signals.
2. **Air feeding system and sensor block:** modeled while taking into account the fluid dynamic aspect and the characteristic times of the sensors (*i.e.*, thermometer, pressure switch and flow switch). The generated data are saved as time series on temporary files.
3. **Valve blocks:** modeled as adjustable restrictions controlled with a Boolean signal taking into account the fluid-dynamic aspect and implementation delays (Fig. 5).
4. **Pipe blocks:** modeled from the fluid dynamics and heat transfer point of view.
5. **Cylinder block:** the fluid-dynamic, mechanical, and thermal parts of the cylinders are modeled. The latter takes into account the velocity damping systems included in the final section of the cylinder and the speed controller valves outside the cylinder. This block also contains the modeling of possible failure types: air leakage is modeled as an adjustable restriction between the two chambers or between the chambers and the outside, and the state of the seal as a parameter that changes the friction force of the plunger. These parameters can be set with configurable constants prior to simulation. (Fig. 4).

After simulation, we read the simulated time series and compare them against the readings of the measuring instruments with the goal of obtaining the same results as the real system. During the reading, the acquisition frequency of the real system, the interaction with the cylinder's limit sensors, and any post-processing are taken into account.

In order to optimize the computing load and the amount of data transmitted, we later do not to use the time series directly. Instead, we represent the operation of the system with only a few particular values. For example, for each pneumatic cylinder, the actuation delay, the time of arrival, the airflow at departure, the airflow at arrival, the maximum airflow, the amount of air absorbed during the movement, the average pressure, and the minimum pressure are collected.

After the model is created, a calibration is performed using the available data, and the values obtained are compared with the actual values to check for a match.

To model wear damage, after an analysis of component failure modes (Nakutis & Kaškonas, 2008; Belforte, Raparelli, & Mazza, 1992; J. Chen, Zio, Li, Zeng, & Bu, 2018), effective parameters are identified to represent the state of the system. In the analyzed use case, the possible leakage is characterized as three adjustable local restrictions (between the first chamber and the environment, between the second chamber and the environment, or between the chambers) while the seal state is an adjustable friction coefficient.

Simulations are performed from an ideal operating state corresponding to the HI of 100% (the state of the part at the time of system calibration, assumed to be healthy) to a failure state, corresponding to the HI of 0%. The law by which the condition is calculated depends on the use that is made

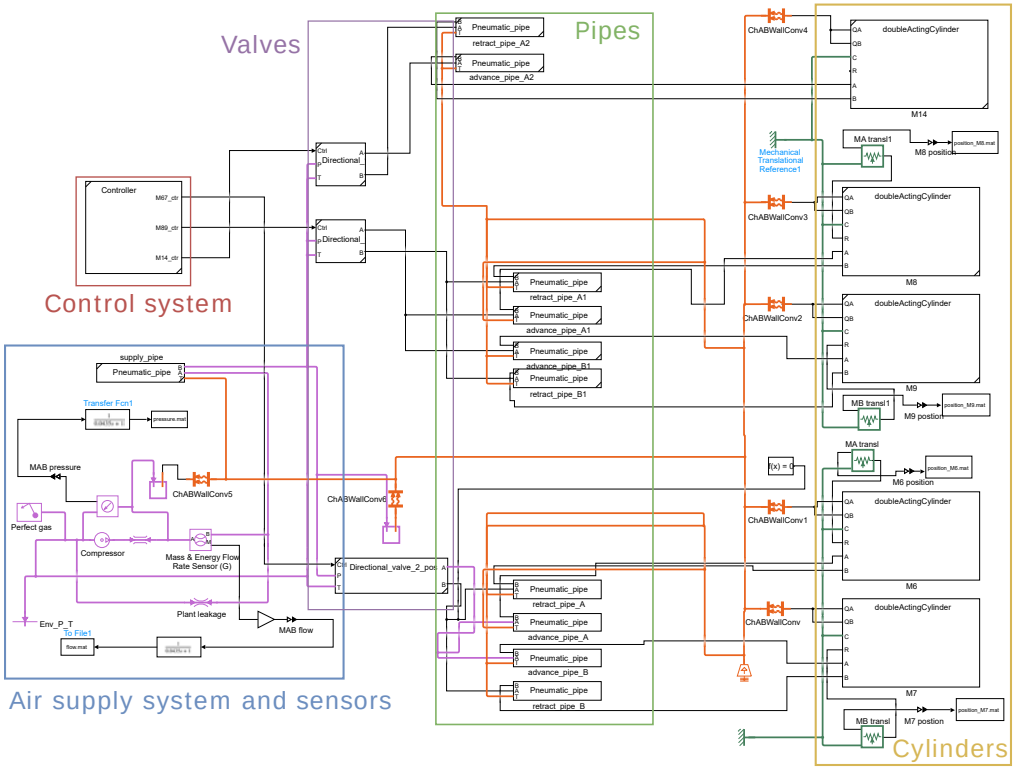


Figure 3. View of the physical model with macro blocks and the structure of major ones highlighted.

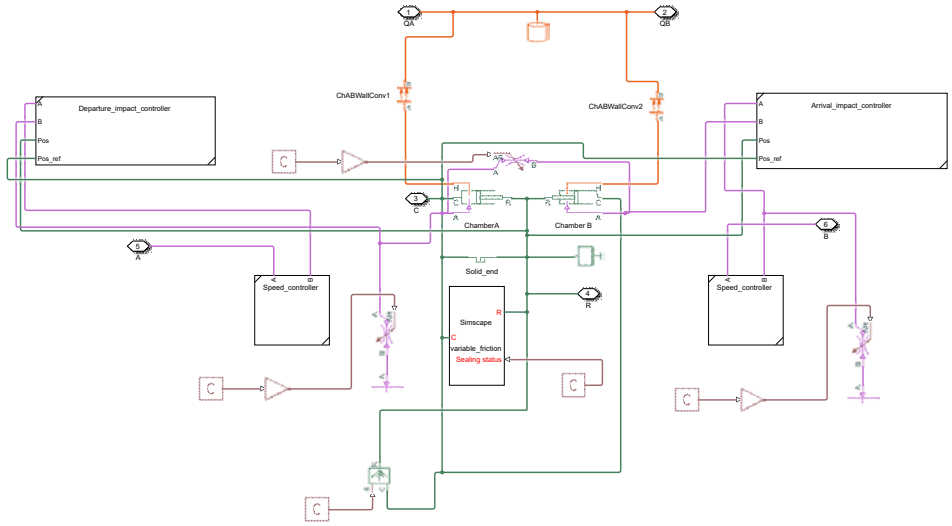


Figure 4. Detail view of one of the blocks modeling the behavior of pneumatic cylinders. It can be seen the variable restrictions that model leakage and the customized block that models seal friction.

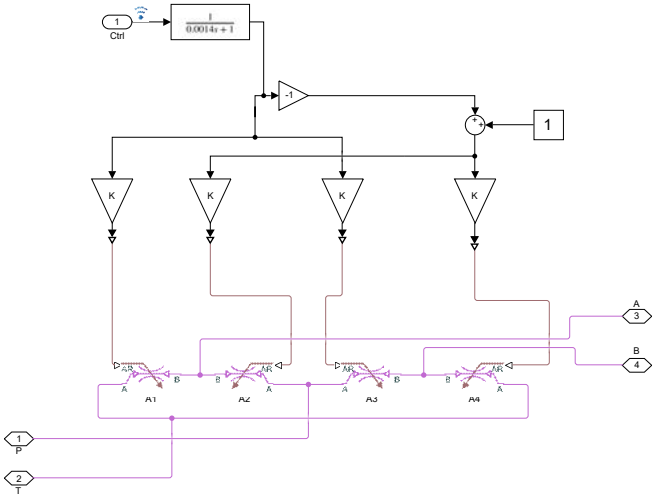


Figure 5. Detail view of one of the blocks modeling the behavior of pneumatic valves.

of the monitored machine and the requirements that it must have. For example, in the case of pneumatic cylinders, it is calculated based on the change in movement time and energy used.

3.2. Parameter perturbation in simulation model

To make the model effective under different operating conditions, we run several simulations by varying the system boundary conditions and the damage progression law. In the logic of keeping the algorithm simple and applicable to different types of models, each failure mode was treated independently (thus each parameter modeling wear progresses at different rates but does not affect others). Of course, the effects that these parameters have on the operation of the simulated device sum up and affect the HI.

This process can be schematized in the following points (see Fig. 6):

1. **Preparation:** The model is calibrated from the conditions measured in the real system. Then it is determined what is the critical value of each of the parameters that represent the damage (the value that alone would bring the part HI to 0) through a series of simulations in which the main operating parameters are varied (see Fig. 7) and from which matrices of critical values are obtained.

2. **Generation of the modeled system:** At this stage, the system's own characteristics, those that are not expected to change over time are determined. These parameters are the damage progression laws for each component and each failure mode, and also additional static parameters such as the position of the control valves. Every damage progression is determined by a quadratic law defined as follows:

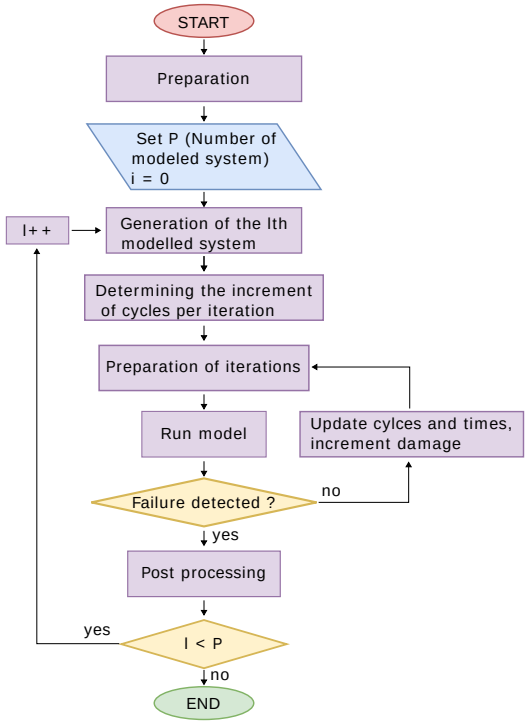


Figure 6. The flowchart of the script running the simulations

$$\begin{cases} (n < n_f): & Kd_{i,j}(n) = n^2 \cdot \frac{(2-4\cdot nli)}{n_f^2} + n \cdot \frac{4\cdot nli-1}{n_f} \\ (n \geq n_f): & Kd_{i,j}(n) = \frac{2n_f(2nli-1)-x(4nli-3)}{n_f} \end{cases} \tag{1}$$

model where $Kd_{i,j}$ is the damage progression coefficient for the failure mode i of the device j, n is the number of cycles made by the device, $nli \in [0.25, 0.7]$ is the *non-linearity index*, n_f is the number of cycle that, would bring the part to failure. The law is formulated to be always increasing while being parabolic up to the n_f value, and then proceeds linearly. The damage progression coefficient multiplied by the critical damage values determines the value of the corresponding failure parameter.

3. **Determining the increment of cycles per iteration:** For each model defined in the previous step, a series of simulations must be run in which the number of cycles performed by the machine increases progressively so as to increase the various damage parameters. The simulations must proceed until the HI of at least one of the parts drops to 0. To keep the computation time acceptable, it was decided that only one cycle is actually modeled in each simulation while the cycle counter and time are recalculated using these formulas:

$$n_k = n_{k-1} + min\left(n_{f,i,j}\right)/N \tag{2}$$

$$t_k = t_{k-1} + min\left(n_{f,i,j}\right) \cdot t_n \tag{3}$$

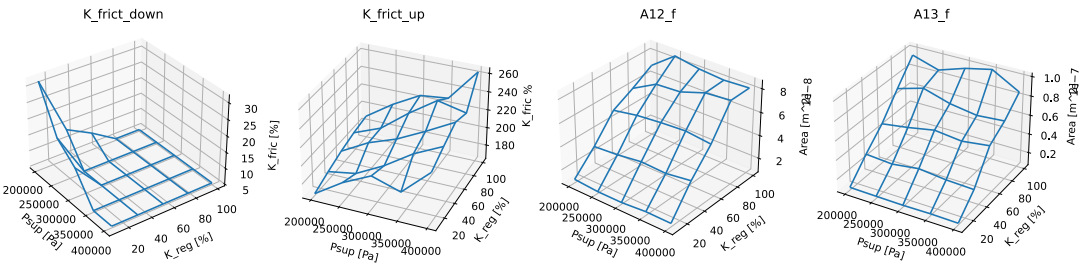


Figure 7. The figure shows the critical matrices for a pneumatic cylinder. The matrices represent the critical value of various damage parameters (friction reduction, friction increase, leakage between chambers, leakage between the second chamber and the environment) as a function of supply pressure and speed control valve adjustment. It can be seen that the critical friction is greatly influenced by pressure while the critical value of leakage is more influenced by valve adjustment.

Where, n_k in the number of cycles made by the device in the iteration k, $n_{f,i,j}$ is the cycles to fail for the failure mode i of the device j, N is the desired number of iterations, t_k is the virtual time in which the simulation k takes place, and t_n is the number of cycles performed in the unit of time.
This reduces the computational time by simulating only a number of cycles equal to about N, which are equally spaced. In practice, the number of simulations may vary in relation to N, as the overlapping effects of increasing different damage parameters may dampen or accentuate their impact on the part condition.

4. **Preparation of iteration:** Before each simulation, some parameters that may vary during the lifetime of the part, such as environmental parameters, are randomly calculated. Specifically, in the presented use case, the pressure of the air supply system, the ambient pressure, the ambient temperature, the sampling start time, the pressure drop of the supply system, and the friction coefficient of the sealing of each cylinder are varied.
5. **Run to failure:** For each iteration first the damage parameters are calculated using the equation:

$$P_{i,j,k} = Kd_{i,j,k}\left(n_k\right) \cdot P_{crit,i,j} \tag{4}$$

where $P_{i,j,k}$ is the damage parameter for the failure mode i of the device j and the iteration k, $Kd_{i,j,k}\left(n_k\right)$ is the damage coefficient for the failure mode i of the device j and the iteration k (see Eq. 1 and 2), $P_{crit,i,j}$ is the critical parameter value for the failure mode i of the device j computed from the critical matrix crested in the preparation phase.
After the computation of the damage parameters, the physical model is run, and finally, the HI of each part is calculated. In case one of the parts has a HI equal to or less than zero the iterative process is terminated otherwise the next one is run.
6. **Post-processing:** At the end of the iterations, the collected data is saved and labeled with the RUL and the HI (depending on the use case) related to the corresponding iteration.

4. Data analysis and prediction model performance

In the conducted studies we assume that labelled data from the target domain (physical system) is available, which categorizes the proposed method as a supervised DA solution (Motiian, Piccirilli, Adjeroh, & Doretto, 2017). In fact, as soon as the generated R2F data from the simulation and the labelled asset sensor readings are available, the acquired data can be used to train a prediction model. In addition, gradient boosted trees (T. Chen & Guestrin, 2016) have been used to estimate the health status of the physical system, defined as HI or RUL, given the sensor readings from it. For evaluating the performance boost from PDP, two separate regression models (XGBR) with the same complexity will be trained using the following datasets:

- Limited annotated data (10% of the available data) from the physical system (the model trained with this dataset is referred to as **XGBR1**)
- Limited annotated data (10% of the available data) from the physical system and the R2F from the simulation model by employing PDP (the model trained with this dataset is referred to as **XGBR2**)

4.1. Use cases and experiments

In this industrial project, two different systems, from SMC Schweiz AG[1] and TCI engineering[2], are tested as use cases to inspect the scalability of the proposed method. The former, is a pneumatic pick and place demonstrator which can be used to mimic different failure types given various working conditions (*e.g.*, by changes in the main pressure of the compressed air). For this use case, the deployed model is used to predict the RUL of the system as degradation in the physical system does not cause any significant financial loss. In fact, this demonstrator is used to create real R2F data for testing the accuracy of the RUL predictions.

The latter; however, is used in a production line owned by

[1]https://www.smc.eu/de-ch
[2]https://www.tci-sa.ch/en/

a third-party company and thus no failures can be artificially built during production. For this use case only the HI of the system is predicted as no failure samples are available from the physical system. As shown later, given the changes in the working conditions, the customer's needs and minor inspections, there are numerous fluctuations in the HI values, resulting them not to be monotonic.

In what follows, the results of model performance comparison for predicting the RUL and HI is provided in detail. However, the similar results of feature importance distribution for RUL and HI predictions, subsection 4.3, and the robustness to the noise, subsection 4.5, are not included in order to conserve space. Nonetheless, this omission does not diminish the completeness of this paper in any manner.

4.2. Model accuracy for predicting the HI

The XGBR1 and XGBR2 prediction models are tested on the data attained from the physical system which were not previously exposed to them during model training. These model have the R^2 scores of 0.676 and 0.853 respectively, which indicates the superiority of the model trained with simulation data. Moreover, for ensuring that sample selection does not impact the accuracy comparison between XGBR1 and XGBR2, these accuracy values are calculate as the mean of accuracy given different seeds for sampling data from the physical system.

In addition, Fig. 8 demonstrates the progression of the HI of the physical system during approximately 3 months. Given the fact that, the data acquisition from the physical system did not start right after a maintenance, there is no reference asset behavior which represent a fully healthy behavior. Therefore, the calculated HI values are equally biased to start from a value which is as close to 1 as possible. In addition, as stated in subsection 4.1, numerous internal and external factors constantly impact the physical system during production, which prevents a monotonic HI sequence.

4.3. Feature importance distribution

The trained XGBRs can provide insight about the importance of different sensor readings from the system. In fact, importance values indicate how influential one feature is in determining the output of the prediction model. Considering that, data reading from industrial assets is not always perfect, it can be expected that there are scenarios where the attained data from an asset contains noise, has missing values or in an extreme case the installed sensor does not provide any data. In such situations, it is vital to determine the role of different sensor readings and also try to train models which use a wider range of sensor readings from the physical system. By doing so, erroneous predictions from the model are prevented, resulting in enhanced model reliability.

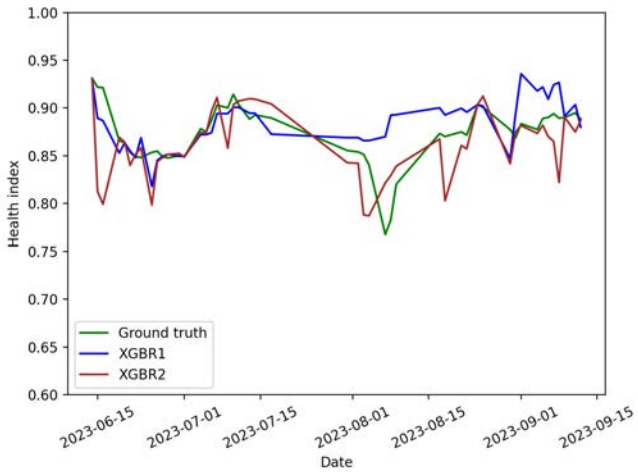


Figure 8. Asset HI along with XGBR1 and XGBR2 predictions

On a separate note, by comparing the significance of different asset readings, it is also possible to verify if the trained prediction model has converged to a PdM solution or is merely a preventive maintenance model relying on the cycle number.

As it can be seen in figures 9 and 10, given the sparsity of data from the physical model, the XGBR trained only using the data from the physical system has a highly unbalanced feature importance across different readings of the studied asset. Therefore, from a model reliability point of view, the simulation model can significantly enhance the performance of the model. Please note that for sake of anonymity and data protection for the involved industrial partner in the conducted studies, hashed sensor reading names are provided in the aforementioned figures.

4.4. Model comparison for predicting the RUL

The XGBR1 and XGBR2 trained for the SMC Schweiz AG demonstrator have fairly similar R^2 scores of 0.980 and 0.917 respectively. The higher accuracy in predicting the RUL is partially attributed to fact that the controlled laboratory setup allows us to control all the boundary conditions. Whereas in the shopfloor, the system is influenced by numerous factors that cannot be controlled or predicted, *e.g.*, a drop in pressure or regulation intervention. Additionally, it is also due to the fact that the induced failure in the system is a result of (semi-)linear opening of the valves in the demonstrator for mimicking leakage in various parts of the system. Nonetheless, as discussed in subsection 4.3, the XGBR1 model puts more emphasis on the information about the cycle count and ignores the information from the rest of the available sensor readings making it a less desirable solution given PdM requirements, especially in case that a faster failure than the ones seen before during model training occurs in the system.

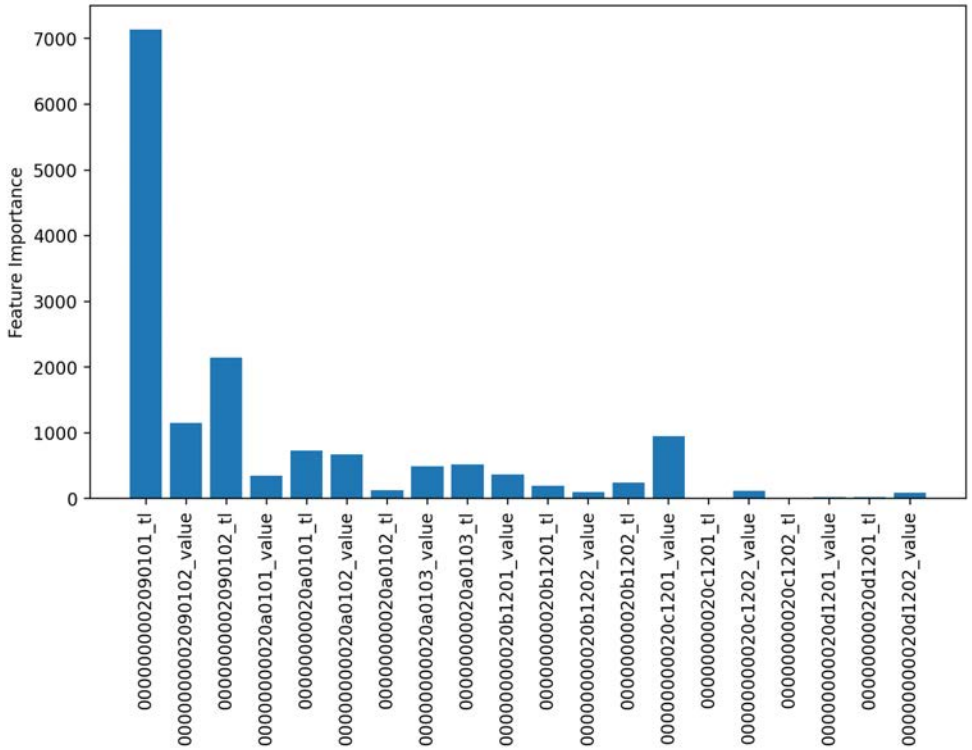


Figure 9. Feature importance among different asset sensor readings for XGBR1 model. This model relies on a limited number of sensor readings from the system

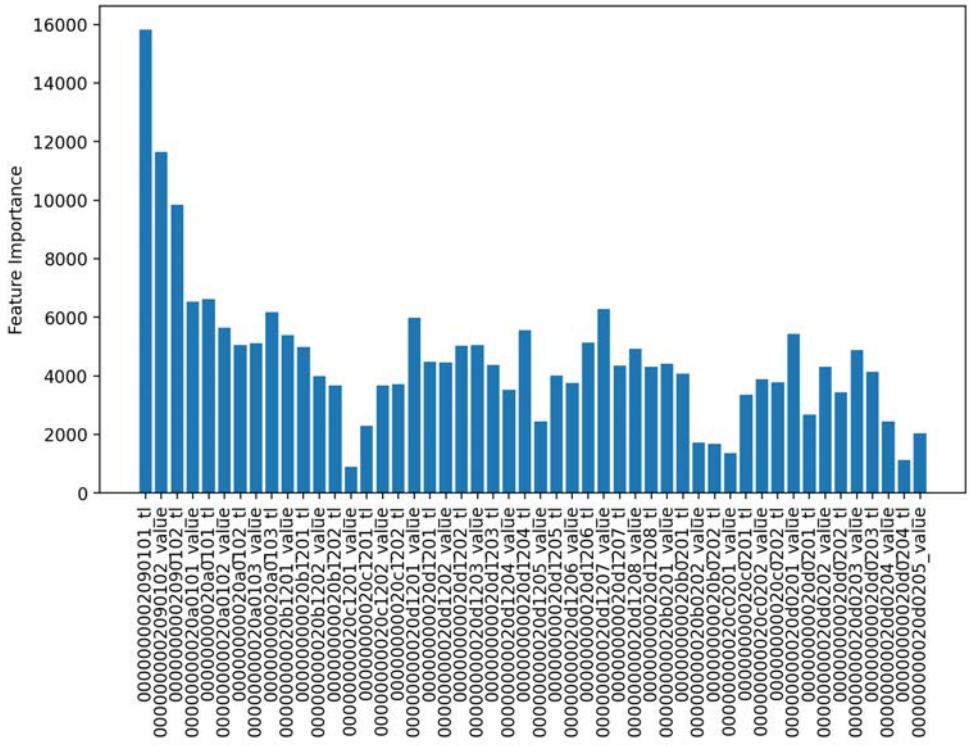


Figure 10. Feature importance among different asset sensor readings for XGBR2 model. This model has a more spread-out feature importance across different sensor readings of the system

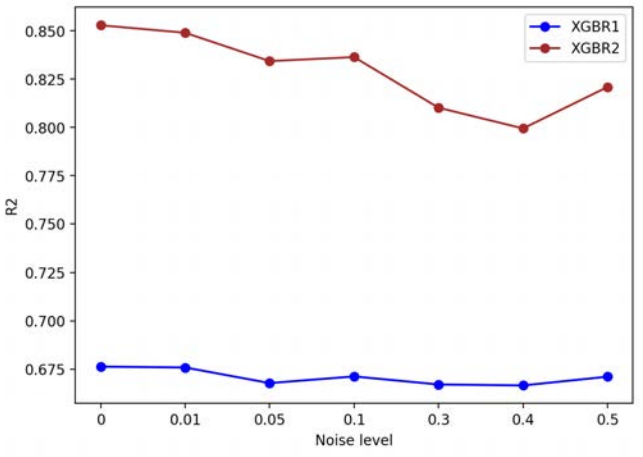


Figure 11. Impact of asset reading noise on the accuracy of the trained prediction models

4.5. Robustness to asset sensor reading noise

In this part of the conducted studies, the impact of noise on the most decisive asset readings on the accuracy of the XGBR1 and XGBR2 for the TCI engineering case are examined. For manipulating the asset sensor readings, 8 of the most influential features given the values in figures 9 and 10 are selected. Thereafter, samples of each of these features ($x^i \in X$) are distorted as follows:

$$x^i_{new} = x^i + Max\{x^j\}_{j=1}^{|X|} \times noise\,level \times d \sim N(0,1) \quad (5)$$

where $|X|$ is the total number of feature readings, $noise\ level$ is scalar value (see Fig 11) and d is a sampled value from $N(0,1)$ which represents normal distribution with mean of 0 and standard deviation of 1. Fig 11 shows the impact of noise on the R^2 score of the predictors. As it can be seen, regardless of the added noise value, the performance of the model trained with the additional simulation data is superior which suggest the robustness of the trained model compared to the prediction model trained only with scarce data from the physical system.

5. Discussion and Conclusion

Gathering annotated data from a physical system for developing PdM solutions is one of the most time-consuming and expensive steps which inhibits many end users in industry for utilizing the full potential of their production assets. In the conducted studies, we aimed to introduce a novel approach for RUL and HI prediction model training which uses data generated from a simulation model and a minimal set of samples from the physical system as apposed to complete R2F datasets from the asset. We aimed to highlight the importance of simulation data generation with PDP for covering as much as of the parameter space of an asset for enhancing the performance of the prediction model despite the scarcity of asset readings. It was shown how the proposed method, in the best case scenario, increases the R^2 score of the trained model by 26% while simultaneously using a wider range of sensor readings from the physical system. Furthermore, the results of model performance deterioration in presence of asset reading noise demonstrated that, regardless of the added noise to the readings, the R^2 score of the model trained with the additional simulation model data is higher. For the future work, we aim to develop classifiers for different working conditions of an asset and then find the corresponding regions of the data generated from the simulation model for a more fine-tuned data generation. In addition, we aim to inspect the performance of the prediction model in different regions of the parameter space and generate data from the simulation model which explicitly can boost the performance of the model in the chosen region of the parameter space.

References

Belforte, G., Raparelli, T., & Mazza, L. (1992). Analysis of typical component failure situations for pneumatic cylinders under load. *Lubrication engineering*, *48*(11), 840–845.

Bonomi, N., Cardoso, F., Confalonieri, M., Daniele, F., Ferrario, A., Foletti, M., ... Pedrazzoli, P. (2021). Smart quality control powered by machine learning algorithms. *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 764–770.

Chang, Y., Fang, H., & Zhang, Y. (2017). A new hybrid method for the prediction of the remaining useful life of a lithium-ion battery. *Applied energy*, *206*, 1564–1578.

Chen, D., Meng, J., Huang, H., Wu, J., Liu, P., Lu, J., & Liu, T. (2022). An empirical-data hybrid driven approach for remaining useful life prediction of lithium-ion batteries considering capacity diving. *Energy*, *245*, 123222.

Chen, J., Zio, E., Li, J., Zeng, Z., & Bu, C. (2018). Accelerated Life Test for Reliability Evaluation of Pneumatic Cylinders. *IEEE Access*, *6*, 75062–75075. (Conference Name: IEEE Access)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cortes, C., & Mohri, M. (2011). Domain adaptation in regression. In *International conference on algorithmic learning theory* (pp. 308–323).

Cui, L., Du, S., & Hawkes, A. G. (2012). A study on a single-unit repairable system with state aggregations. *IIE Transactions*, *44*(11), 1022–1032.

Didona, D., & Romano, P. (2014). On bootstrapping machine learning performance predictors via analytical models.

arXiv preprint arXiv:1410.5102.

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.

Fathi, K., Sadurski, M., Kleinert, T., & van de Venn, H. W. (2023). Source component shift detection classification for improved remaining useful life estimation in alarm-based predictive maintenance. In *2023 23rd international conference on control, automation and systems (iccas)* (p. 975-980). doi: 10.23919/ICCAS59377.2023.10316874

Fathi, K., van de Venn, H. W., & Honegger, M. (2021). Predictive maintenance: an autoencoder anomaly-based approach for a 3 dof delta robot. *Sensors*, *21*(21), 6979.

Fei, C. (2022). *Lithium-ion battery data set.* IEEE Dataport. Retrieved from https://dx.doi.org/10.21227/fh1g-8k11 doi: 10.21227/fh1g-8k11

Ferrario, A., Confalonieri, M., Barni, A., Izzo, G., Landolfi, G., & Pedrazzoli, P. (2019). A Multipurpose Small-Scale Smart Factory For Educational And Research Activities. *Procedia Manufacturing*, *38*, 663–670.

Gao, Y., Liu, X., Huang, H., & Xiang, J. (2021). A hybrid of fem simulations and generative adversarial networks to classify faults in rotor-bearing systems. *ISA transactions*, *108*, 356–366.

Hanachi, H., Liu, J., Banerjee, A., Chen, Y., & Koul, A. (2014). A physics-based modeling approach for performance monitoring in gas turbine engines. *IEEE Transactions on Reliability*, *64*(1), 197–205.

Kaufmann, E., Loquercio, A., Ranftl, R., Müller, M., Koltun, V., & Scaramuzza, D. (2020). Deep drone acrobatics. *arXiv preprint arXiv:2006.05768*.

Lei, Y., Li, N., Gontarz, S., Lin, J., Radkowski, S., & Dybala, J. (2016). A model-based method for remaining useful life prediction of machinery. *IEEE Transactions on reliability*, *65*(3), 1314–1326.

Lin, R., Yu, Y., Wang, H., Che, C., & Ni, X. (2022). Remaining useful life prediction in prognostics using multi-scale sequence and long short-term memory network. *Journal of computational science*, *57*, 101508.

Liu, J., Wang, W., Ma, F., Yang, Y., & Yang, C. (2012). A data-model-fusion prognostic framework for dynamic system state forecasting. *Engineering Applications of Artificial Intelligence*, *25*(4), 814–823.

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the ieee international conference on computer vision* (pp. 5715–5725).

Nakutis, Žilvinas., & Kaškonas, P. (2008). An approach to pneumatic cylinder on-line conditions monitoring. *Mechanics*, *72*(4), 41–47.

Nejjar, I., Geissmann, F., Zhao, M., Taal, C., & Fink, O. (2024). Domain adaptation via alignment of operation profile for remaining useful lifetime prediction. *Reliability Engineering & System Safety*, *242*, 109718.

Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In *2018 ieee international conference on robotics and automation (icra)* (pp. 3803–3810).

Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., & Li, X. (2020). Adversarial transfer learning for machine remaining useful life prediction. In *2020 ieee international conference on prognostics and health management (icphm)* (pp. 1–7).

Rahat, M., Mashhadi, P. S., Nowaczyk, S., Rognvaldsson, T., Taheri, A., & Abbasi, A. (2022). Domain adaptation in predicting turbocharger failures using vehicle's sensor measurements. In *Phm society european conference* (Vol. 7, pp. 432–439).

Saxena, A. (2023). *Nasa turbofan jet engine data set.* IEEE Dataport. Retrieved from https://dx.doi.org/10.21227/pjh5-p424 doi: 10.21227/pjh5-p424

Si, X.-S., Wang, W., Hu, C.-H., Zhou, D.-H., & Pecht, M. G. (2012). Remaining useful life estimation based on a nonlinear diffusion degradation process. *IEEE Transactions on reliability*, *61*(1), 50–67.

Thelen, A., Li, M., Hu, C., Bekyarova, E., Kalinin, S., & Sanghadasa, M. (2022). Augmented model-based framework for battery remaining useful life prediction. *Applied Energy*, *324*, 119624.

Tiboni, G., Arndt, K., & Kyrki, V. (2023). Dropo: Sim-to-real transfer with offline domain randomization. *Robotics and Autonomous Systems*, *166*, 104432.

Wang, Q., Taal, C., & Fink, O. (2021). Integrating expert knowledge with domain adaptation for unsupervised fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–12.

Yang, B., Lei, Y., Jia, F., & Xing, S. (2019). An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mechanical Systems and Signal Processing*, *122*, 692–706.

Yu, K., Fu, Q., Ma, H., Lin, T. R., & Li, X. (2021). Simulation data driven weakly supervised adversarial domain adaptation approach for intelligent cross-machine fault diagnosis. *Structural Health Monitoring*, *20*(4), 2182–2198.

Dynamic Modeling of Distributed Wear-Like Faults in Spur Gears: Simplified Approach with Experimental Validation

Lior Bachar[1], Roee Cohen[1], Omri Matania[1], and Jacob Bortman[1]

[1]*Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, 8410501, Israel*

liorbac@post.bgu.ac.il
coroe@post.bgu.ac.il
omrimat@post.bgu.ac.il
jacbort@bgu.ac.il

ABSTRACT

Dynamic models of gears are recognized for offering a promising platform for gaining a profound understanding of the dynamic response, particularly the vibration signature. Wear is considered among the most common and concerning fault mechanisms in gears, yet its recognition and subsequent diagnosis remain challenging. In this study, we introduce an existing dynamic model of spur gear vibrations and extend its validation for distributed wear-like faults. The novelty of this work lies in addressing the complexities associated with modeling distributed faults using simplified yet sophisticated approaches. These involve variance among defected teeth, calculation of time-variant gear mesh stiffness, and consideration of the forces induced by the fault. The model is validated through pioneering controlled experiments, analyzing dozens of degrading distributed wear-like faults. This comparison verifies our capability to generate realistic simulations of vibration signals from worn gears. To bridge the discrepancy between the induced and simulated faults, the model first constructs the healthy profile of the inspected gear, incorporating manufacturing errors and tooth modifications. Subsequently, meticulous photography enables the replication of faults in the model with a high resemblance to the experiment. The results demonstrate a strong correlation between measured and simulated signals, as verified through trend analysis of features extracted from synchronous average signals in both the cycle and order domains. This study lays the groundwork for in-depth investigation into the physics of gear wear, paving the way for potential applications such as fault severity estimation and intelligent fault diagnosis in future studies.

1. INTRODUCTION

Predictive maintenance of gear wear is crucial, considering gears' pivotal role in rotating machinery and their constant susceptibility to failure due to operation in harsh regimes. Gear fault types can be broadly classified into localized faults such as breakage and cracks, and distributed faults such as abrasive wear and fatigue pitting. Abrasive wear, caused by oil contamination and sliding motion, leads to continuous destruction of the tooth surface, posing a viable risk of catastrophic failure due to reduced gear efficiency and high stress concentrations. Nevertheless, Feng, Ji, Ni, and Beer (2023) reviewed the latest developments in gear wear monitoring and demonstrated that diagnosing gear wear through vibration analysis is still challenging due to the intricate patterns manifested in the signature that remain unresolved. Physical models, such as tribological models (Archard, 1953) and dynamic models (Liang, Zuo, and Feng, 2018; Mohammed & Rantatalo, 2020), have been suggested over the years in order to bridge this gap.

Most of the published dynamic models of gear wear typically analyze the effects of wear on the time variant gear mesh stiffness (gms). Liu, Yang, and Zhang (2016) utilize a spur gear model to study the changes in the gms and transmission error due to wear, as well as the wear expression in the vibrations. Brethee, Zhen, Gu, and Ball (2017) introduce a helical gear model, validated through endurance tests, and analyze both the gms and the increase in frictional excitation with wear. Many other studies (Chen, Lei, and Hou, 2021; Cui et al., 2023; Ren & Yuan, 2022; Shen et al., 2020), incorporate Archard's tribological model in the dynamic model to calculate the worn surface, demonstrating the effect of wear on the dynamic characteristics. However, most of these models lack experimental validation, and in general, the coefficients in Archard's equation are largely unknown, making their evaluation in the models nontrivial.

In this work, we introduce a novel approach for modeling distributed wear-like faults in spur gears, validated through

experimentation. By leveraging simplifying assumptions, this modeling approach effectively captures the dynamic response of worn gears, while also facilitating its adaptation by other scholars in their models. In Section 2, we introduce the framework of the existing dynamic model, establishing the groundwork for this study. Section 3 delves into the simplified wear modeling approach, while Section 4 presents the experimental setup. Model validation is detailed in Section 5, accomplished through vibration analysis of the synchronous average signals and their spectrum. Finally, Section 6 concludes this work, providing insights and suggesting potential directions for future research.

2. DYNAMIC MODEL

This study adopts the dynamic model for spur gears proposed by Dadon, Koren, Klein, and Bortman (2018), which has been experimentally validated for healthy gears and various localized faults, serving as the foundation for this study. The simulated system has an open gearbox with torsional shafts connecting the driving pinion to a motor and the driven gear to a brake applying external torque, as illustrated in Figure 1. The vector of generalized coordinates (u) consists of 13 degrees of freedom: six for each wheel, representing linear displacement (x_i,y_i,z_i) and angular position (θ_i,φ_i,ψ_i) in space (where i=p,g), and another for the brake's angle (θ_b). Figure 2 presents a block diagram illustrating the model's stages. The vibration signal is generated by solving the Euler-Lagrange equations of motion, as described in Eq1.

$$M\ddot{u} + C\dot{u} + K(u) \cdot u = F(t, u) \quad (1)$$

Here, M, C, K, and F are the mass, damping, and stiffness matrices, and the excitation force vector, respectively. The non-linearity in $K(u)$ arises from the time-variant gear mesh stiffness (gms), computed using the potential energy method. The model is configured with parameters such as gear module, number of teeth, tooth width, and surface quality, alongside operational conditions like input speed and load. In contrast to many published models, where the gms is the sole non-linear component, this model introduces non-linearity in the excitation force vector. It incorporates deviations from the involute profile, such as surface roughness and faults, as displacement inputs along the Line of Action (LoA), which are subsequently transformed into forces by appropriately multiplying them with the gms. Thus, the excitation force consists of three components overall: the motor torque, the brake torque, and the force induced by displacements along the LoA, as shown in Eq. 2.

$$F = k_{\theta_p}\theta_m \cdot \hat{\theta}_p + T_b \cdot \hat{\theta}_b + gms \cdot \delta \cdot \bar{c} \quad (2)$$

Here, k_{θ_p} is the input shaft torsional stiffness, θ_m is the motor's angle, T_b is the brake's torque, δ is the displacement along the LoA, and $\bar{c}$ is a vector of geometric coefficients projecting this force onto u. $\hat{\theta}_p$ and $\hat{\theta}_b$ are unit vectors pointing to θ_p, θ_b, respectively.

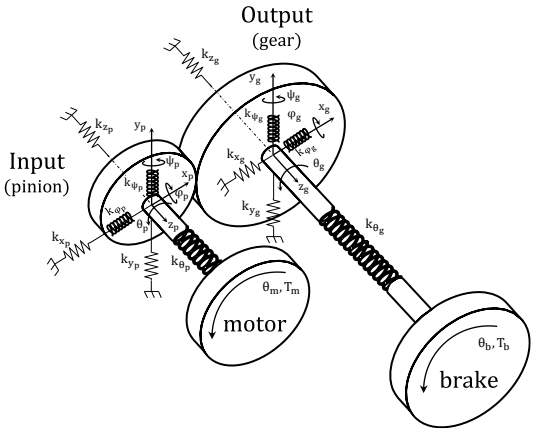


Figure 1. The simulated system (Dadon et al., 2018).

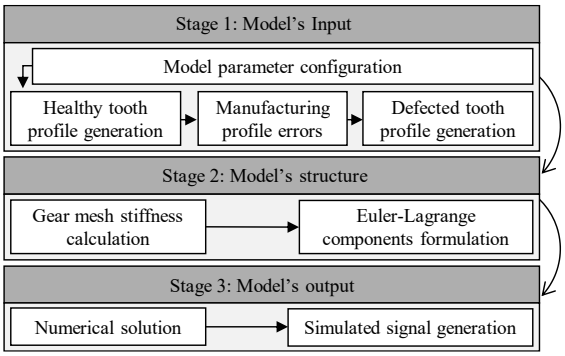


Figure 2. A schematic block diagram of the dynamic model.

3. GEAR WEAR MODELING

Any model development is grounded in premises that aim to balance the tradeoff between simplicity and reality (Mohammed & Rantatalo, 2020). We make the following assumptions that simplify our ability to simulate gear wear:

Assumption I: The worn profile is linear (or piecewise linear) and uniform along the tooth width, as illustrated in Figure 3.

Assumption II: The nominal parameters of the worn profile vary slightly among teeth.

Assumption III: The worn profile influences the cross-section properties of the tooth, thereby impacting the potential strain energy and, consequently, the gms.

Assumption IV: Contact properties such as contact ratio, pressure angle, and initial contact point remain unchanged. However, any deviation from the nominal LoA is treated as a displacement input in the excitation force.

It is crucial to acknowledge the limitations of these assumptions, as the gear wear mechanism is more complex, involving details not covered by these simplifications, such as sliding motion and improper contact. Nonetheless, these simplifications establish a foundation for comprehending the general wear behavior. The following subsections explore the effects of wear on the gms and the excitation force.

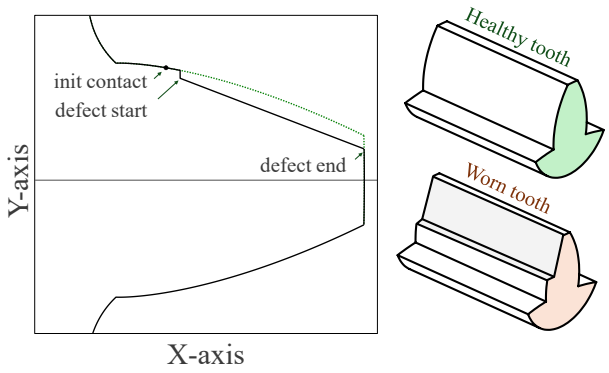


Figure 3. The simulated worn profile.

3.1. Effects of Wear on the Gear Mesh Stiffness

The computation of the gms commonly involves two steps, both are affected by wear. The first step employs the potential strain energy method for calculating the equivalent stiffness of a meshing tooth pair from engagement to separation. In this case, the influence of wear is self-evident, as tooth geometry is changed, and potential strain energy is derived from integration with respect to volume. The second step involves combining the equivalent stiffness of all tooth pairs based on the contact ratio governing the transition from a single pair to a double pair. In a healthy state, the equivalent stiffness can be computed once and then replicated and concatenated to form the cyclic gms. This procedure stays largely similar in case of localized faults, except for swapping the equivalent stiffness of one healthy pair with that of the defected pair. However, with distributed wear faults, where the worn profile varies among teeth, the equivalent stiffness is calculated individually for each tooth pair, and then meticulously combined, as depicted in Figure 4.

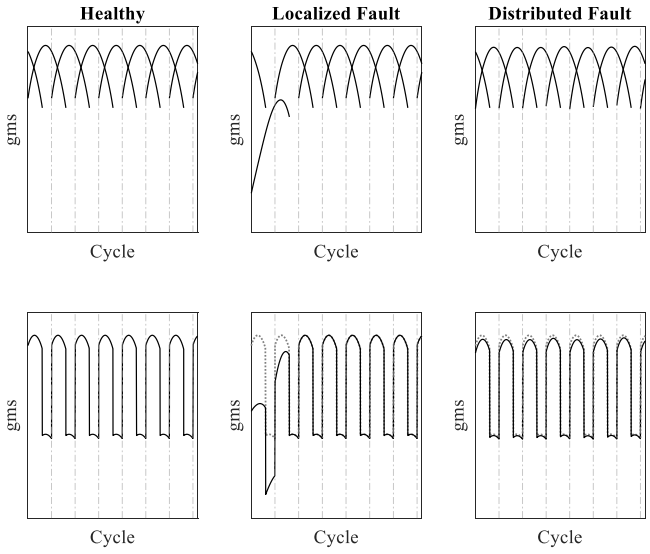


Figure 4. Construction of the gms signal in case of a healthy status, localized fault, and distributed fault.

3.2. Effects of Wear on the Excitation Force

One of the non-trivial assumptions made is that contact properties remain wear-invariant. While this assumption may be controversial, it is not without basis when appropriately compensated. As explained previously, deviations from the LoA are treated as displacement inputs in the excitation forces. The computation of the fault displacement involves straightforward geometric manipulations according to Eq. 3, using parameters depicted in Figure 5. Given that the fault displacement is unique for each tooth, it is multiplied by the equivalent stiffness of its corresponding pair, and the resulting product is then combined using the same procedure as in the gms, as depicted in Figure 6.

$$\delta_{\text{fault}} = (Y_{\text{invlt}} - Y_{\text{def}}) \cdot \frac{\cos(\gamma)}{\cos(\varphi + \gamma)} \tag{3}$$

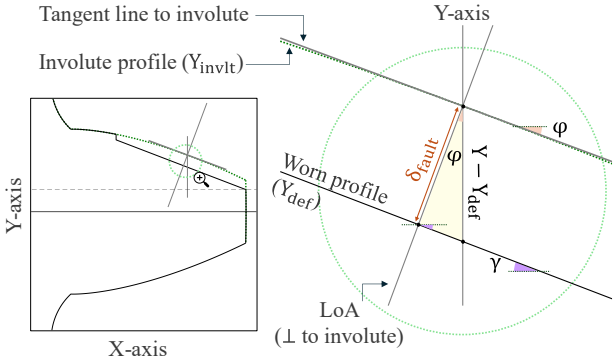


Figure 5. An illustration of the fault displacement calculation and the required parameters.

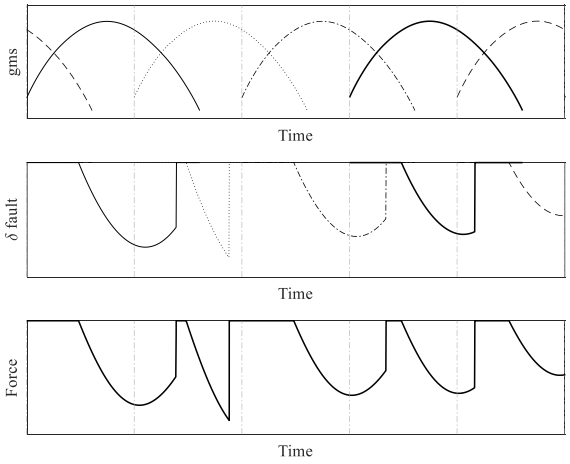


Figure 6. Generation of excitation force through fault displacement along the LoA, multiplied by the gms.

4. Experimental Setup

We conducted an extensive controlled experiment for model validation, employing a dedicated test apparatus for spur gears. Vibration data were collected for both a healthy (H) status and 35 degrading wear cases (W_i), using piezoelectric

accelerometers, alongside rotational speed measured by tachometers, as depicted in Figure 7. Details regarding the experimental program and gearbox parameters can be found in Table 1. The degradation of a reference tooth throughout the experiment is showcased in Figure 8. This figure includes photographs illustrating three cases corresponding to the beginning, middle and the end of the experiment. Additionally, a heatmap depicts the contour of all wear cases, with the color gradient correlating with fault severity.

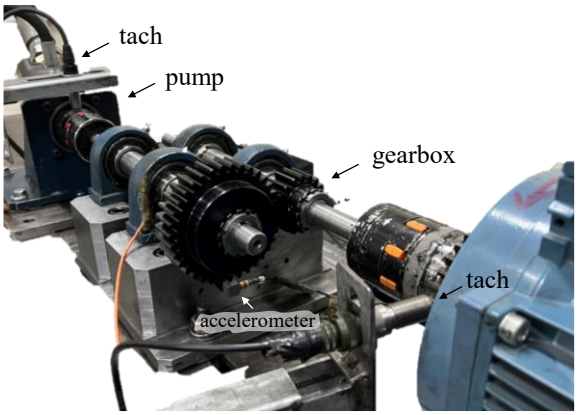


Figure 7. The experimental setup employed for validation.

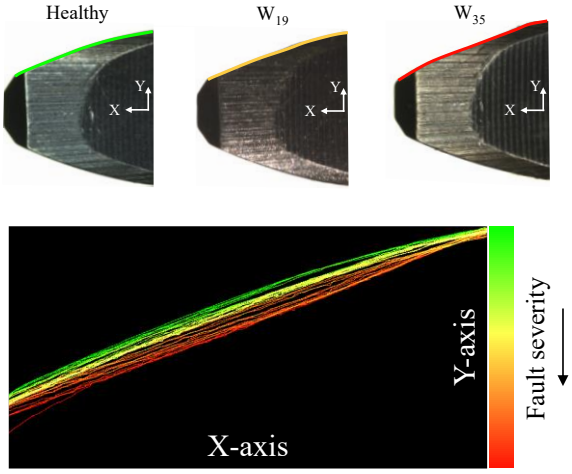


Figure 8. Photographs and heatmap depicting the profile degradation of a reference tooth throughout the experiment.

Table 1. Gearbox parameters and experimental program.

Gearbox parameters	
Module	3mm
Reduction ratio	35:18
Precision grade	DIN8
Experimental program	
Input speed	15rps, 45rps
Output Load	10Nm
Sampling rate	50kS/s
Signal duration	60s
Health status	H, $\{Wi\}_{i=1}^{35}$

5. Model Validation

The validation of the proposed wear modeling approach is empirical, relying on a qualitative comparison between simulation and experiment. This comparison involves analyzing trends in energy-based (such as rms) and shape-based (such as kurtosis) features extracted from the synchronous average (SA) signal and the difference signal in the cycle domain, and the SA spectrum in the order domain (Matania, Bachar, Bechhoefer, and Bortman, 2024). For both simulated and measured data, the SA is computed after the raw vibration signal undergoes angular resampling based on the output shaft's speed. It is essential to note that while the simulated signal is directly calculated at the wheels' center, the measured signal is significantly influenced by the transmission path between the gearbox and the sensor. This influence results in expected differences in spectral behavior, such as attenuations and resonances (Bachar et al., 2021, 2023). Consequently, experimental and simulated results are presented with left and right y-axes, with energy-based features normalized by the healthy (H) status according to Eq. 4, ensuring comparability of general trends.

$$F_{norm} = \frac{|F - F_H|}{F_H} \tag{4}$$

5.1. SA Analysis in the Cycle Domain

Figure 9 compares SA signals at 45rps in healthy status and for severe wear. Expected impulses appear in all the SAs. Both experiment and simulation show an amplified signal without sharp and rare impulsive responses due to wear, as expected. This observation highlights challenges in wear monitoring, as faulty signals may not emphasize the fault, creating a false impression of a healthy transmission.

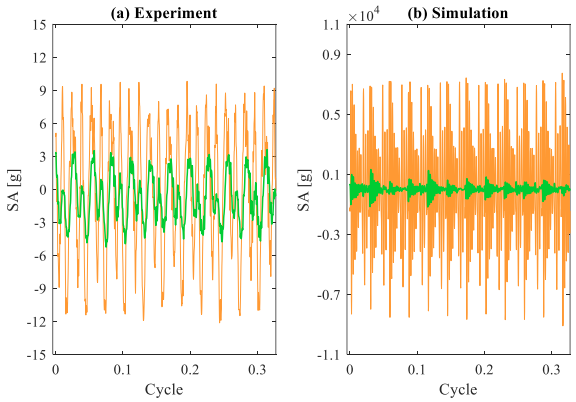


Figure 9. Comparison of SA signals between simulation (right) and experiment (left) at 45rps, in healthy (green) and severe wear (orange) statuses.

Figure 10 analyzes trends in SA rms, difference rms, and difference kurtosis across fault severity. Experimental results are depicted with error bars representing the scattering in the feature values. The following insights can be derived from these results:

- There is a strong correlation between simulation and experiment in rms trends, particularly evident in difference rms, where rms increases monotonously with wear degradation. Moreover, the "wavy" trend is observable in both simulation and experiment, suggesting that this behavior may have a physical basis.
- The higher speed (45rps) exhibits superior correlation between simulation and experiment. The purportedly weaker correlation at 15rps may be attributed more to the effects of speed and transmission path on the vibration signature (Bachar et al., 2021) rather than discrepancies.
- Kurtosis values are generally low and remain relatively stable. Given that kurtosis emphasizes sharp, rare impulses, which distributed wear faults are not expected to, it might not be suitable for gear wear monitoring. This insight is evident both in simulation and experiment.
- In most cases, discrepancies between simulation and experiment are more evident in more severe faults.

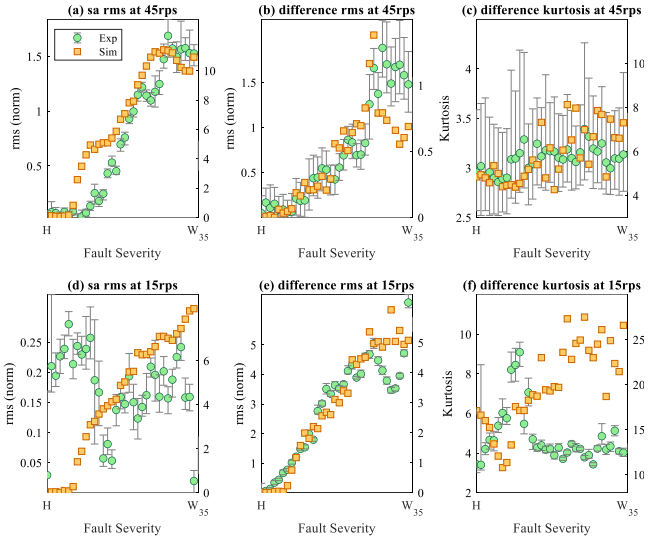


Figure 10. Trend analysis of features extracted from the SA and difference signals across fault severity at 45rps (top row) and 15rps (bottom row).

5.2. SA Analysis in the Order Domain

A comparison of the SA spectra across fault severity is presented in the spectrograms in Figure 11. For clarity, the top row in each spectrogram, corresponding to the healthy status (H), is thicker and separated from the degrading wear cases by a line. High amplitudes at the gearmesh harmonics are observed in all spectra, as expected. Furthermore, across both speeds and for both simulation and experiment, the general behavior with respect to fault severity is similar; that is, the spectral energy mostly varies monotonously with health degradation. However, the optimal wear manifestation for fault detection and degradation monitoring varies across different frequency bands for each combination of data source and rotational speed, as expected. The differences in spectrum background are expected to lead to such discrepancies, but as long as they are acknowledged, focus can be placed on the similarities obtained between experiment and simulation.

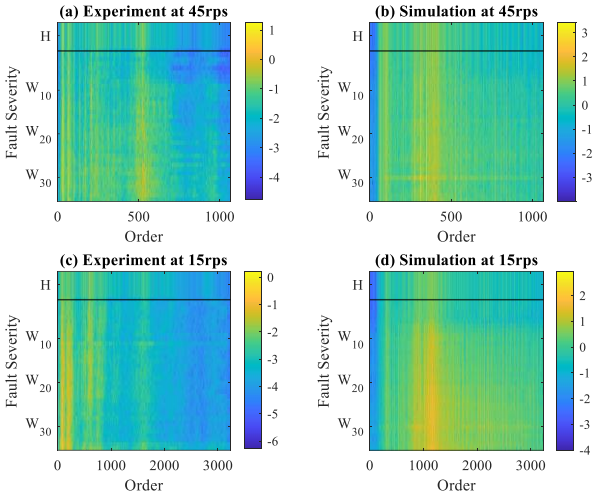


Figure 11. SA spectra across fault severity between experiment (left) and simulation (right) at 45rps (top) and 15rps (bottom).

Early research on gear monitoring (Randall, 1982) demonstrates the impact of distributed wear faults on gear mesh harmonics and modulation sidebands in the spectrum. To capture similar behaviors between simulation and experiment, we compute the gear mesh energy (gme) and modulation sideband energy (sbe) in the spectrum (X) according to Eq. 5-6 and compare their trends across fault severity, as depicted in Figure 12.

$$\text{gme} = \sum_{n=1}^{gm_{max}} |X(gm \times n)| \tag{5}$$

$$\text{sbe} = \sum_{n=1}^{gm_{max}} \sum_{m=1}^{gm/2} |X(gm \times n \pm m)| \tag{6}$$

Here, gm_{max} is the maximum number of gearmesh harmonics available within bandwidth. The following insights can be derived from the spectral analysis results:

- There is a strong correlation between simulation and experiment in both gme and sbe trends, closely mirroring the energy-based feature analysis in the cycle domain in Figure 10.
- Both spectral energies exhibit a monotonic variation with wear degradation, displaying the same "wavy" trend as discussed in the cycle domain analysis.

The spectral analysis aligns with the feature analysis in the cycle domain, confirming the similarity between simulated

and measured signals. Despite relying on a set of non-trivial assumptions, the proposed simplified modeling approach successfully captures the general wear patterns in the simulated signal. Moreover, while features for monitoring localized faults focus mainly on signal shape and modulation sidebands, gear mesh energy is also crucial for diagnosing wear faults. This holds true in both simulation and experiment across different speeds. Discrepancies between simulation and experiment, stemming from the assumptions made, are more pronounced in severe wear faults, as expected due to the challenging assumption of invariant contact properties with wear. Nevertheless, the strong similarities between simulation and experiment validate the model's ability to generate a simulated vibration signal reflecting the fundamental characteristics of gear wear.

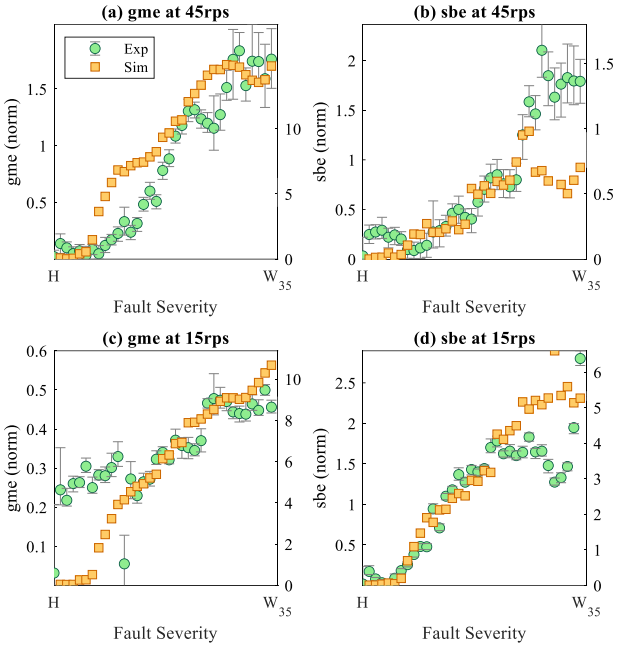


Figure 12. Comparison of spectral analysis of the gme (left) and sbe (right) between simulation and experiment at 45rps (top row) and 15rps (bottom row).

6. Conclusion

Gear wear monitoring is crucial for predictive maintenance, yet identifying patterns in the vibration signature associated with wear remains challenging. This study aims to bridge this gap by introducing a novel, simplified approach for simulating distributed wear-like faults. We make a set of assumptions to investigate wear characteristics essential for health monitoring, incorporating wear faults into an existing framework of dynamic model for gear vibrations. We demonstrated the impact of wear on the non-linear gear mesh stiffness and the excitation force according to the proposed modeling. Extensive controlled experiments validate our approach, comparing experimental and simulated results across different speeds and fault severities. A visual examination of the synchronous average signal and its spectrum across fault severity confirms that the proposed wear modeling closely resembles the experimental results, yielding similar insights. In-depth trend analyses of features in both cycle and order domains reveal crucial insights into the intricacies of wear monitoring, capturing "wavy" trends as the fault deteriorates. This underscores the importance of analyzing energy-based features, such as gear mesh energy and sideband energy, rather than shape-based features, for monitoring distributed wear faults. The strong correlation between experimental and simulated results confirms the feasibility of our approach, suggesting it as a simple yet effective enhancement for simulating wear faults in any standard dynamic gear model. Our study opens avenues for practical applications, including refining the simplified model for real-system applications and developing novel methods for wear prediction in future work.

References

Archard, J. F. (1953). Contact and Rubbing of Flat Surfaces. *Journal of Applied Physics*, *24*(8), 981–988. https://doi.org/10.1063/1.1721448

Bachar, L., Dadon, I., Klein, R., & Bortman, J. (2021). The effects of the operating conditions and tooth fault on gear vibration signature. *Mechanical Systems and Signal Processing*, *154*, 107508. https://doi.org/10.1016/J.YMSSP.2020.107508

Bachar, L., Matania, O., Cohen, R., Klein, R., Lipsett, M. G., & Bortman, J. (2023). A novel hybrid physical AI-based strategy for fault severity estimation in spur gears with zero-shot learning. *Mechanical Systems and Signal Processing*, *204*. https://doi.org/10.1016/j.ymssp.2023.110748

Brethee, K., Zhen, D., Gu, F., Theory, A. B.-M. and M., & 2017, undefined. (2017). Helical gear wear monitoring: Modelling and experimental validation. *ElsevierKF Brethee, D Zhen, F Gu, AD BallMechanism and Machine Theory, 2017•Elsevier*. http://dx.doi.org/10.1016/j.mechmachtheory.2017.07.012

Chen, W., Lei, Y., Fu, Y., theory, L. H.-M. and machine, & 2021, undefined. (2021). A study of effects of tooth surface wear on time-varying mesh stiffness of external spur gear considering wear evolution process. *ElsevierW Chen, Y Lei, Y Fu, L HouMechanism and Machine Theory, 2021•Elsevier*. https://doi.org/10.1016/j.mechmachtheory.2020.104055

Cui, Q., Dong, N., Cui, Q., Tong, J., Wang, R., Lu, H., Dong, N., Zhou, J., Tong, R., Wang, H., & Lu, F. (2023). Study on wear evolution of spur gears considering dynamic meshing stiffness. *Springer*, *37*(7), 2023. https://doi.org/10.1007/s12206-023-0606-3

Dadon, I., Koren, N., Klein, R., & Bortman, J. (2018). A realistic dynamic model for gear fault diagnosis. *Engineering Failure Analysis*, *84*. https://doi.org/10.1016/j.engfailanal.2017.10.012

Feng, K., Ji, J. C., Ni, Q., & Beer, M. (2023). A review of vibration-based gear wear monitoring and prediction techniques. In *Mechanical Systems and Signal Processing* (Vol. 182). https://doi.org/10.1016/j.ymssp.2022.109605

Liang, X., Zuo, M. J., & Feng, Z. (2018). Dynamic modeling of gearbox faults: A review. *Mechanical Systems and Signal Processing*, *98*, 852–876. https://doi.org/10.1016/J.YMSSP.2017.05.024

Liu, X., Yang, Y., International, J. Z.-T., & 2016, undefined. (2016). Investigation on coupling effects between surface wear and dynamics in a spur gear system. *ElsevierX Liu, Y Yang, J ZhangTribology International, 2016•Elsevier*. https://doi.org/10.1016/j.triboint.2016.05.006

Matania, O., Bachar, L., Bechhoefer, E., & Bortman, J. (2024). Signal Processing for the Condition-Based Maintenance of Rotating Machines via Vibration Analysis: A Tutorial. *Sensors 2024, Vol. 24, Page 454*, *24*(2), 454. https://doi.org/10.3390/S24020454

Mohammed, O. D., & Rantatalo, M. (2020). Gear fault models and dynamics-based modelling for gear fault detection – A review. *Engineering Failure Analysis*, *117*, 104798. https://doi.org/10.1016/J.ENGFAILANAL.2020.104798

Randall, R. B. (1982). A New Method of Modeling Gear Faults. *Journal of Mechanical Design*, *104*(2), 259–267. https://doi.org/10.1115/1.3256334

Ren, J., Coatings, H. Y.-, & 2022, undefined. (2022). A dynamic wear prediction model for studying the interactions between surface wear and dynamic response of spur gears. *Mdpi.ComJ Ren, H YuanCoatings, 2022•mdpi.Com*. https://doi.org/10.3390/coatings12091250

Shen, Z., Qiao, B., Yang, L., Luo, W., Yan, R., Manufacturing, X. C.-P., & 2020, undefined. (2020). Dynamic modeling of planetary gear set with tooth surface wear. *ElsevierZ Shen, B Qiao, L Yang, W Luo, R Yan, X ChenProcedia Manufacturing, 2020•Elsevier*. https://doi.org/10.1016/j.promfg.2020.06.010

BIOGRAPHIES

Lior Bachar is currently a current Ph.D. student in mechanical engineering in BGU-PHM LAB at Ben-Gurion University of the Negev, Beer-Sheva, Israel. His primary academic interest is vibration-based diagnosis of gears. His research, published in numerous papers, employs hybrid physical data-driven based approaches, integrating experimentations, signal processing, dynamic modeling, and AI applications. Lior completed with honors his bachelor's degree and master's degree in mechanical engineering in Ben-Gurion University of the Negev.

Roee Cohen is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Roee completed with honors his bachelor's degree and master's degree in mechanical engineering in Ben-Gurion University of the Negev.

Omri Matania is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Omri is a Talpiot graduate and served nine years in IDF in several roles including algorithm section leader. He completed with honors his bachelor's degree in mathematics and physics in the Hebrew University of Jerusalem and completed his master's degree with honors in mechanical engineering in Ben-Gurion University of the Negev

Jacob Bortman is currently a full Professor in the department of mechanical engineering and the head of the PHM Lab in Ben-Gurion University of the Negev. Retired from the Israeli air force as brigadier general after 30 years of service with the last position of the head of material directorate. Chairman and member of several boards: director of business development of Odysight Ltd, Chairman of the board of directors, Selfly Ltd., board member of Augmentum Ltd., board member of Harel finance holdings Ltd., Chairman of the board of directors, Ilumigyn Ltd. Editorial board member of: “Journal of Mechanical Science and Technology Advances (Springer, Quarterly issue)”. Head of the Israeli organization for PHM, IACMM - Israel Association for Comp. Methods in Mechanics, ISIG - Israel Structural Integrity Group, ESIS - European Structural Integrity Society. Received the Israel National Defense prize for leading with IAI strategic development program, Outstanding lecturer in BGU, The Israeli prime minister national prize for excellency and quality in the public service - First place in Israel. Over 80 refereed articles in scientific journals and in international conference.

Enhanced Diagnostics Empowered by Improved Mechanical Vibration Component Extraction in Nonstationary Regimes

Fadi Karkafi[1,2], Jérôme Antoni[1], Quentin Leclère[1], Mahsa Yazdanianasr[3,4], Konstantinos Gryllias[3,4], and Mohammed El Badaoui[2,5]

[1] *INSA Lyon, LVA, UR677, 69621 Villeurbanne, France*

[2] *Safran Tech, Rue des Jeunes Bois – Châteaufort 78772 Magny–les–Hameaux, France*

[3] *KU Leuven, Department of Mechanical Engineering, Celestijnenlaan 300, 3001, Leuven, Belgium*

[4] *Flanders Make @ KU Leuven, Belgium*

[5] *UJM-St-Etienne, LASPI, UR3059, F-42023, Saint-Etienne, France*
fadi.karkafi@insa-lyon.fr
jerome.antoni@insa-lyon.fr
quentin.leclere@insa-lyon.fr
mahsa.yazdanianasr@kuleuven.be
konstantinos.gryllias@kuleuven.be
mohammed.el-badaoui@safrangroup.com

ABSTRACT

When analyzing vibration and sound signals from rotating machinery, accurately tracking individual orders is crucial for diagnostic and prognostic objectives. These orders correspond to sinusoidal components, also known as deterministic signals, whose amplitude and phase are modulated in response to the angular speed of the machine. The extraction of these components leads to a more comprehensive approach to differential diagnostics. When the machine operates under varying conditions, consistently tracking the orders becomes challenging, particularly in nonstationary regimes with very fast variations. Typically, this issue is addressed using common techniques such as Vold-Kalman filter (VKF), where the bandwidth of the selective filter is adjusted to handle the speed variations. However, in the presence of high-speed fluctuations, manual adjustment of these weights becomes difficult to balance the compromise between achieving accurate tracking by effectively filtering around the speed variations, and maintaining a low estimation bias by reducing noisy errors. To overcome this constraint, the proposed methodology is driven by the need to integrate speed fluctuations into an optimal solution using VKF. This adaptation involves the consideration of angular acceleration profiles within the innovation process. In this context, the bandwidths are automatically adjusted to their optimal values according to the machine's regime. Optimality is achieved by crafting a model dependent on the order signal-to-noise ratio (SNR) and the auto-regression coefficient. This optimization allows for a practical adjustment tailored to the distinctive characteristics of each order. A comprehensive analysis of the resulting model transfer function reveals crucial insights into the impact of the given order SNR and the speed fluctuations. Subsequently, the methodology undergoes performance assessment through simulations and synthetic cases, showcasing its viability and effectiveness across various regimes. Notably, its practical application is highlighted in envelope-based bearing diagnosis, during operations characterized by variable-speed conditions, thus underlining its promise in real-world applications.

1. INTRODUCTION

In mechanical systems, gears and rolling-element bearings play vital roles in power transmission, necessitating reliable operation (Randall & Antoni, 2011). Vibration and acoustic signals emitted by these mechanical pieces (Braun, 1986) exhibit distinct cyclostationary (CS) behavior (Antoni, 2009). Gears generate first-order cyclostationary (CS1) components with a periodic mean, forming a harmonic spectrum corre-

sponding to their fundamental period. Rolling-element bearings, on the other hand, exhibit second-order cyclostationary (CS2) components marked by periodic autocovariance and an instantaneous envelope. These distinctions are crucial for differential diagnosis (Antoni & Randall, 2002), providing insights into overall system health. In stationary conditions, Fourier analysis simplifies the identification of CS1 components, which are characterized by their corresponding amplitude and phase. This facilitated the rise of order tracking, enabling the extraction of these characteristic components. These orders correspond to sinusoidal signals, also termed deterministic components, wherein its amplitude and phase modulations are influenced by the machine angular speed with respect to a reference angle. Subtracting the tracked ones reveals residual random aspects, including the CS2 environment, offering a comprehensive diagnostic perspective.

However, as mechanical systems venture into varying nonstationary regimes, the complexity deepens. These distortions are mainly introduced by the change of the machine power intake and the effect of transmission from the excitation source to the sensor. Even with the application of angular resampling to address frequency modulations stemming from nonstationarity (McFadden, 1989; Bonnardot et al., 2005; Borghesani et al., 2012), the Fourier coefficients representing an order in the spectrum lose their pristine sparsity. This degradation arises from the dynamic evolution of the envelope and is highly emphasized in the presence of high speed fluctuations.

An essential technique in navigating this complexity is the Vold-Kalman filter (VKF) (Vold et al., 1997). Proposed by Vold and Leuridan, VKF has been a cornerstone in the estimation of individual order components instantaneous amplitude and phase. This selective filtering is adapted by adjusting its bandwidth correspondingly to handle nonstationary conditions for each order (M.-C. Pan & Wu, 2007b). Further advancements have led to schemes that simultaneously estimate multiple orders, such as the angular-displacement (AD) (M.-C. Pan & Wu, 2007a) and angular-velocity (AV) (M. Pan et al., 2016) VKF. Considering other notable methods, the Sliding Window Tracking (SWT) (Pai & Palazotto, 2009) technique was introduced to track the varying instantaneous amplitude of a noise-contaminated signal with a moving average. It employs a constant and a pair of windowed regular harmonics to fit the data, providing implicit noise filtering capabilities. Recently, the so-called local synchronous fitting (LSF) has been proposed (Abboud et al., 2022). It was introduced as an enhancement to the global one (GSF) (Daher et al., 2010) in the sense that it estimates a cyclic-nonstationary (CNS) mean, but through a local polynomial fit, using the well-known Savitsky-Golay filter (Savitzky & Golay, 1964). The properties of the filter was also discussed in (Abboud et al., 2019). Interested readers can refer to (Randall et al., 2011) for a comparison among more relevant separation techniques.

While SWT, LSF and VKF are highly accurate in terms of envelope estimation, their limitations become pronounced when mechanical systems transition into fast nonstationary variations. On the one hand, even though SWT attempts to address nonstationary behavior, it assumes a fixed sliding window length, which is not optimal for handling high speed fluctuations. On the other hand, LSF suffers from the estimation of the Fourier coefficient from the centered signal by applying a linear angle-invariant convolution. In fact, from a signal point of view, this operation aims at estimating the mean (trend) of a non-stationary time series. While the classical low-pass filtering is efficient when the noise is (angle-stationary), it can be highly compromised in the case of non-stationary noise, in particular when the noise is impulsive. Finally, while the VKF's bandwidth can be customized to handle nonstationary speed variations, the compromise between accurate tracking and maintaining low estimation bias is emphasized. In addition, within high speed variations, manually adapting the filter nonstationary parameters may pose challenges in achieving optimized solutions leading to unbalance this compromise. Other variants and extended versions of the filter considered tracking the components of interest by maximizing the kurtosis (Dion et al., 2013) and tuning the bandwidth accordingly (Feng et al., 2022) to take into account the high amplitude fluctuation. The high-speed environment necessitates a methodology that explicitly incorporates speed fluctuations into its definition. Consequently, rather than adjusting the filter bandwidth parameters to meet predefined objectives, an optimal approach involves integrating speed fluctuations directly into the model.

Therefore, the paper aims to present a novel VKF optimized solution where the innovation process is directly affected by the speed fluctuations, thus the bandwidth is automatically adapted, yielding stationary hyper-parameters to be tuned: the order signal-to-noise ratio (SNR) and the auto-regression coefficient. Section 2 states the problem with a particular attention to formulating the transmission path effect from a CNS view, followed by a detailed exposition of the methodology in Section 3. Sections 4 and 5 validate the effectiveness of the proposed solution on numerical and experimental signals. In light of the obtained results, the paper is sealed with a general conclusion in Section 6.

2. Problem Statement

This section introduces the basic model that will serve to define the solution, which consists of Fourier series whose complex exponentials are function of the variable angle, θ, being a reference angle in the machine, and the coefficients are only dependent on the speed $\omega = \frac{d\theta}{dt}$, with the load effect omitted. Practically, when vibration and acoustic responses are measured from rotating and reciprocating machinery, the effects of flow noise, turbulence, and transient events are captured, in addition to the sum of rotational dynamics $x(t)$, such that

the total measured signal $y(t)$ is:

$$y(t) = x(t) + \nu(t), \tag{1}$$

where $\nu(t)$ is causal and uncorrelated with $x(t)$ such that it doesn't affect its generation or behavior. A temporal representation of a rotating machine's excitation is defined as

$$x(t) = \sum_k a_k(t) e^{j2\pi\alpha_k\theta(t)+\Phi_k}, \tag{2}$$

where the harmonic cyclic order $\alpha_k = \frac{k}{\Theta}$ such that Θ stands for the angular period of the rotating component of reference, $\theta(t)$ is the angular position of the reference expressed in [rad] and a_k and Φ_k are respectively the slowly varying complex envelopes and phases. In the particular case of constant operating speed (i.e. $\theta(t) = \omega_0 t$), $a_k(t)$ become essentially constant over time, representing the Fourier coefficients in the harmonic series. In such scenario, the synchronous average (SA) is one of the most used tools to extract such components (Braun, 1975; McFadden, 1987), with minimum disruption in the residual signal (Randall et al., 2011). The SA simply consists in slicing the signal (often after an angular resampling step) into cycles equal to the rotation period of the mechanical piece of interest and performing an empirical average to reject (or reduce) non-synchronous components including noise and interfering components. However, in the case of speed varying excitation, the complex envelopes $a_k(t)$ become slowly varying and principally dependent on the operating speed and its fluctuations (Abboud et al., 2016). Despite mild conditions under which higher derivative orders can be neglected, high-speed fluctuations still induce non-stationary behavior in the envelope, potentially leading to inaccurate estimations for each given order. Therefore, the study focuses on the first derivative order to showcase the impact of speed fluctuations on the deterministic component evolution, given by:

$$x(t) = \sum_k a_k(\omega(t)) e^{j2\pi\alpha_k\theta(t)+\Phi_k}. \tag{3}$$

3. Proposed Methodology

The mechanical vibration nature specifies that the envelope functions should be smooth and slowly varying over time. One way of specifying this, is to demand that a repeated difference should be small, which satisfies the following VKF equation in stationary mode,

$$\frac{\partial^q a_k(t)}{\partial t^q} = \varepsilon_k(t), \tag{4}$$

where q is the derivative order and ε_k is a process of uncertainties that degrades the envelope. During this study and for simplicity, an order $q = 1$ will be elaborated. This leads to the fact that, in stationary regime, the envelope will tend to be constant with stationary uncertainties. However, this model turns out to be more sophisticated in the case where $\omega(t)$ is varying with respect to time because the uncertainties also become nonstationary. This can be formulated as follows:

$$\frac{\partial a_k(\omega)}{\partial \omega} = \varepsilon_k(\omega). \tag{5}$$

With the angular velocity varying over time, the application of the chain rule leads to the formulation:

$$\frac{\partial a_k(\omega(t))}{\partial t} = \dot{\omega}(t)\varepsilon_k(\omega(t)). \tag{6}$$

Given the existence of multiple regime scenarios, it is essential to consider both stationary and nonstationary modes, allowing for a generalized modeling approach. This results in envelope uncertainties attributed to

$$\frac{\partial a_k(\omega(t))}{\partial t} = (1 + \lambda\dot{\omega}(t))\varepsilon_k(\omega(t)), \tag{7}$$

where λ serves as a weighting coefficient. To optimize the model based on diverse domain processes derived from (1) and (7), a discrete-time realization is established. The main processes are expressed as follows:

$$\begin{cases} y_k[n] = a_k[n]e^{j\alpha_k\theta[n]} + \nu[n] \\ a_k[n] - \beta_k a_k[n-1] = (1 + \lambda\dot{\omega}[n])\varepsilon_k[n] \end{cases} \tag{8}$$

where $y_k[n]$ represents the raw noisy component corresponding to the k^{th} harmonic and β_k is the auto-regression coefficient whose value is close to 1. In the following, both $\nu[n]$ and $\varepsilon_k[n]$ are supposed to follow complex Gaussian normal distributions with respective variance σ_ν^2 and $\sigma_{\varepsilon_k}^2$: $\nu[n]$ ~ $CN(0, \sigma_\nu^2)$ and $\varepsilon_k[n]$ ~ $CN(0, \sigma_{\varepsilon_k}^2)$. Since $\dot{\omega}[n]$ and $\varepsilon_k[n]$ are independent, one can deduce that their product variance is equal to $(1+\lambda\dot{\omega}[n])^2 \times \sigma_{\varepsilon_k}^2$. Thus, the definition of the two processes:

$$P_1 : \mathbf{Y}_k - \mathbf{A}_k\mathbf{E}_k = \mathbf{V}, \tag{9}$$

where $\mathbf{Y}_k$, $\mathbf{E}_k$ and $\mathbf{V}$ are expressed as follows:

$$\mathbf{Y}_k = \begin{bmatrix} \vdots \\ y_k[n] \\ \vdots \end{bmatrix}, \mathbf{E}_k = \begin{bmatrix} \ddots & & \\ & e^{j\alpha_k\theta[n]} & \\ & & \ddots \end{bmatrix}, \mathbf{V} = \begin{bmatrix} \vdots \\ \nu[n] \\ \vdots \end{bmatrix}.$$

$$P_2 : \mathbf{D}_k\mathbf{A}_k = \mathbf{\Psi}\xi_k, \tag{10}$$

such that $\mathbf{A}_k$ and the sparse matrix $\mathbf{D}_k$ are expressed as

$$\mathbf{A}_k = \begin{bmatrix} \vdots \\ a_k[n] \\ \vdots \end{bmatrix}, \mathbf{D}_k = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\beta_k & 1 & 0 & 0 & 0 \\ 0 & -\beta_k & 1 & 0 & 0 \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & -\beta_k & 1 \end{bmatrix},$$

along with $\mathbf{\Psi}$ and $\boldsymbol{\xi}_k$ expressed as:

$$\mathbf{\Psi} = \begin{bmatrix} \vdots \\ 1 + \lambda\dot{\omega}[n] \\ \vdots \end{bmatrix}, \boldsymbol{\xi}_k = \begin{bmatrix} \vdots \\ \varepsilon_k[n] \\ \vdots \end{bmatrix}.$$

3.1. A Posterior Estimate of the Envelope

With regards to the envelope's dependencies, one can explicitly describe its likelihood and prior distribution. As a result, the a posterior estimation of A_k can be defined by the Maximum a posteriori estimate (MAP):

$$\begin{aligned} \hat{\mathbf{A}}_k &= \text{Argmax}(P(\mathbf{A}_k|\mathbf{Y}_k)) \\ &\propto \text{Argmax}(P(\mathbf{Y}_k|\mathbf{A}_k)P(\mathbf{A}_k)). \end{aligned} \tag{11}$$

After setting $\mathbf{\Gamma}_k = \mathbf{D}_k\mathbf{A}_k$ to simplify the process formulation, the Bayes Theorem can be applied to P_2 in order to compute $P(\mathbf{A}_k)$ as follows:

$$P(\mathbf{A}_k) = \int P(\mathbf{A}_k, \mathbf{\Gamma}_k) d\mathbf{\Gamma}_k. \tag{12}$$

One can deduce from Eqs (10) and (12) that

$$P(\mathbf{A}_k) = P(\mathbf{\Gamma}_k)|\mathbf{D}_k|, \tag{13}$$

To initiate the MAP estimate, the likelihood conditional probability $P(\mathbf{Y}_k|\mathbf{A}_k)$ and the prior one $P(\mathbf{A}_k)$ can be expressed accordingly:

$$P(\mathbf{Y}_k|\mathbf{A}_k) = \frac{1}{\pi\sigma_\nu^{2N}} e^{-\frac{\|\mathbf{Y}_k - \mathbf{E}_k\mathbf{A}_k\|^2}{\sigma_\nu^2}}, \tag{14}$$

$$P(\mathbf{A}_k) = \frac{1}{\pi\sigma_{\varepsilon_k}^{2N}\mathbf{\Omega}} e^{-\frac{\|\mathbf{A}_k^T\mathbf{D}_k^T\mathbf{\Omega}^{-1}\mathbf{A}_k\mathbf{D}_k\|}{\sigma_{\varepsilon_k}^2}} |\mathbf{D}_k|, \tag{15}$$

where $[\mathbf{\Omega}]_{ij} = \delta_{ij}\mathbf{\Psi}[i]^2$ and δ_{ij} is the Kronecker delta. Plugging Eqs (14) and (15) into Eq (11) result minimizing an objective function $J(\mathbf{A}_k)$ represented by:

$$J(\mathbf{A}_k) \approx \frac{\|\mathbf{Y}_k - \mathbf{E}_k\mathbf{A}_k\|^2}{\sigma_\nu^2} + \mathbf{A}_k^T\mathbf{D}_k^T\frac{\mathbf{\Omega}^{-1}}{\sigma_{\varepsilon_k}^2}\mathbf{D}_k\mathbf{A}_k \tag{16}$$

such that $\mathbf{M}^T$ represents the transpose of $\mathbf{M}$. In order to find the minimum, the derivative of $J(\mathbf{A}_k)$ with respect to the variable of interest $\mathbf{A}_k$ is set to 0: $\frac{\partial J(\mathbf{A}_k)}{\partial \mathbf{A}_k} = 0$. As a result, the estimated $\hat{\mathbf{A}}_k$ can be computed as

$$\hat{\mathbf{A}}_k = \left(\mathbb{I} + \frac{\sigma_\nu^2}{\sigma_{\varepsilon_k}^2}\mathbf{D}_k^T\mathbf{\Omega}^{-1}\mathbf{D}_k\right)^{-1}\mathbf{E}_\mathbf{k}^\mathbf{H}\mathbf{Y}_k \tag{17}$$

such that $\mathbf{E}_k^H$ is the conjugate of $\mathbf{E}_k$ and $\mathbb{I}$ is the identity matrix. It can be identified from the resulted model that, $\hat{\mathbf{A}}_k$ depends on the constant ratio $\rho_k = \frac{\sigma_\nu^2}{\sigma_{\varepsilon_k}^2}$ which is none other than the inverse SNR of the k^{th} harmonic.

3.2. Filter Frequency Response Function

In order to interpret the model frequency response function (FRF), a simple scheme is considered in Fig. 1 such that the objective is to find $|H_k(f)|^2$ that would result in $A_k(f)$ from $Y_k(f)$. In order to define $|H_k(f)|^2$, the speed $\omega(t)$ is assumed

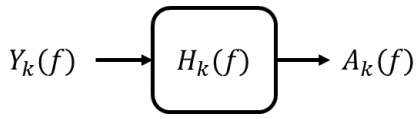


Figure 1. Scheme illustrating the transfer function $H_k(f)$ processing input $Y_k(f)$ to produce output $A_k(f)$.

to be constant or slowly varying, ensuring fixed frequency and allowing the FRF to make sense for each frequency, which results in $\dot{\omega}[n] \approx 0$. In addition, it is noteworthy that, as the filter automatically adapts to speed fluctuations, its stationary parameters remain valid regardless of the operating regime. Consequently, the FRF will be analyzed primarily under stationary conditions, ensuring correct parameters estimation. Recalling the processes described in Eq. (8), its respective power spectral densities would take the following form:

$$S_{Y_k}(f) = E|Y_k(f)|^2 = E|A_k(f)|^2 + E|V(f)|^2, \tag{18}$$

$$S_{A_k}(f) = E|A_k(f)|^2 = \frac{E|\xi_k(f)|^2}{|1 - \beta_k e^{-j2\pi f}|^2}. \tag{19}$$

Knowing that $E|V(f)|^2 = \sigma_\nu^2$ and $E|\xi(f)|^2 = \sigma_\varepsilon^2$, S_{H_k} can be computed as follows:

$$S_{H_k}(f) = \frac{S_{A_k}(f)}{S_{Y_k}(f)} = \frac{1}{1 + \rho_k|1 - \beta_k e^{-j2\pi f}|^2}. \tag{20}$$

Based on Eq. (20), it seems to behave like a low pass filter. For a comprehensive understanding, it is crucial to investigate the parameters of the FRF and the frequencies associated with typical filters. In this analysis, the following parameters are defined in the logarithmic:

$$\begin{cases} \ln(|H(0)|^2) = -\mu_0 \\ \ln(|H(f_p)|^2) = -\mu_p \\ \ln(|H(f_s)|^2) = -\mu_s \end{cases} \tag{21}$$

such that

1. f_p: The passband frequency.
2. f_s: The stopband frequency.
3. $-\mu_0$: Log of the FRF at $f = 0$.
4. $-\mu_p$: Log of the FRF at the passband frequency f_p.
5. $-\mu_s$: Log of the FRF at the stopband frequency f_s.

To illustrate these parameters, consider Figure 2 that depicts an example of lowpass filter with the passband frequency f_p, stopband frequency f_s and other associated parameters.

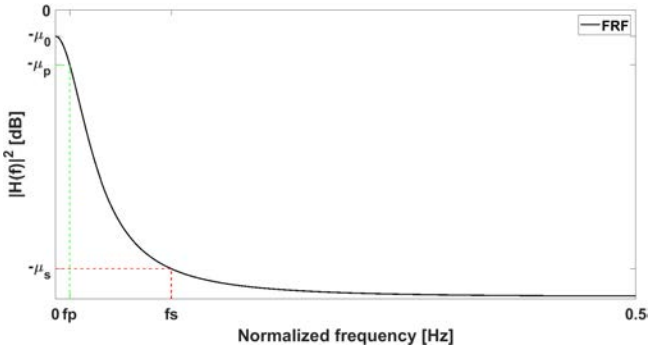
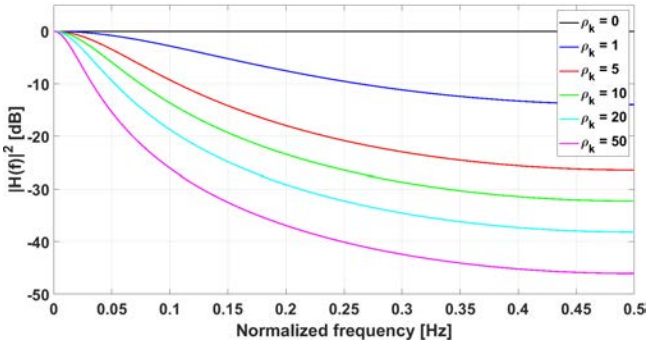
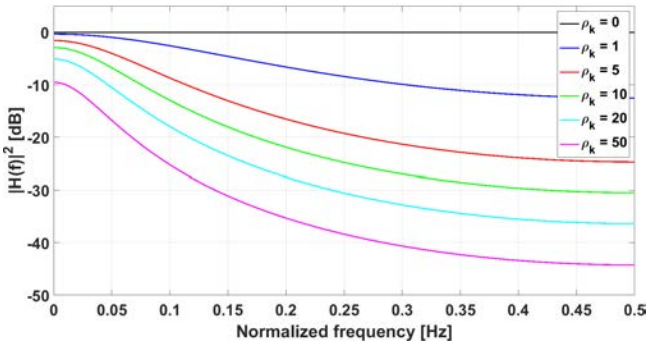


Figure 2. Illustration of a theoretical first-order lowpass filter with the corresponding characteristics.

In Figures 3a and 3b, the FRF for the proposed filter is illustrated under various conditions. Each figure depicts the FRF for different values of ρ_k, showcasing the impact of this parameter on the filter's behavior.



(a)



(b)

Figure 3. Plots of the frequency response function using various ρ_k values with respect to (a) $\beta_k = 1$ and (b) $\beta_k = 0.8$.

Furthermore, the illustrations explore the influence of β_k, presenting results with $\beta_k = 1$ & $\beta_k = 0.8$ respectively. The distinctive curves in each plot reveal the sensitivity of the filter to the model parameters ρ_k and β_k, providing valuable insights into its performance characteristics.

4. Numerical Experiment

Deterministic components encountered in rotating machine signals operating under nonstationary regimes are generally sinusoidal components whose amplitudes and phases are modulated through smooth functions. These amplitude modulations typically arise from resonance encounters or shifts in internal forces. Phase alterations may stem from variations in time delays resulting from mechanical system transfer functions interacting with excitation frequencies, or from torsional oscillations in shafts. To capture these dynamics, the following sinusoidal model is adopted:

$$y[n] = \sum_{k=1}^{3} a_k[n] sin\left(2\pi k \sum_{m=0}^{n} \omega[m] + \Phi_k[n]\right) + \nu[n] \tag{22}$$

where

1. $y[n]$ is the generated signal sampled at 10kHz over a 10 second acquisition window (i.e. resulting in 100 ksamples)
2. $\omega[n]$ stands for the instantaneous frequency of the reference rotating shaft (i.e. the fundamental frequency of the process associated with the order 1) simulated using a first order autoregressive process (see top Figure 4),
3. $a_k[n]$ and $\Phi_k[n]$ are respectively the amplitude and the phase modulations associated with the k^{th} harmonic (see middle and bottom Figure 4), made of linear combination of the square of $\omega[n]$, $\nu[n]$ is a stationary gaussian noise.

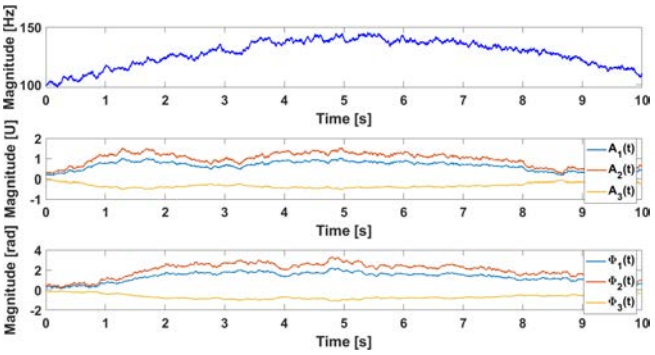


Figure 4. Plots of: (top) the speed constructed using a first order auto-regressive process, (middle) the 3 amplitudes $a_k[n]$ and (bottom) the 3 phase modulations $\Phi_k[n]$ associated with the sinusoids of the synthetic signal.

The signal-to-noise (power) ratio equals -10dB. The proposed methodology is implemented on the signal, alongside three other methods for comparative analysis: SWT, LSF, and VKF with stationary bandwidth in order to assess the efficacy of the speed adaptation proposed in this study. For SWT, the window length was optimized to achieve the best performance,

resulting in a length of 125 samples per sliding window. LSF is applied to the signal after resampling in the angular domain, employing a polynomial order of 5 and a window length of 75 samples. The conventional and proposed VKF parameters, namely β_k and ρ_k, are fine-tuned to achieve the values presented in Table 1. These parameters are crucial as they directly impact the filtering process and subsequent signal analysis. The actual (noise-free) signal, the estimated signals and corresponding errors are displayed in Figure 5 for each method. It is clear from the figure that the proposed estimation is more accurate than the other used techniques: Table 2 shows that the proposed error signal has a significantly lower relative mean error compared to SWT, LSF and VKF.

Table 1. Filter tuned parameters for the simulation case.

Parameters	Value
β_k	$\{0.988, 0.991, 0.957\}$
ρ_k	$\{45, 35, 75\}$
λ	1

Table 2. Relative mean errors of each used method.

Used method	Relative mean error %
SWT	4.18
LSF	8.44
VKF	6.60
Proposed method	**1.84**

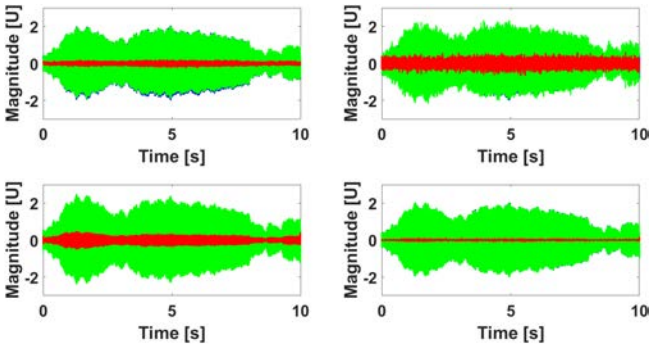


Figure 5. The actual noise-free signal (blue line), the deterministic signal estimate (green line) and the corresponding errors (red line) using the (top-left) SWT method, (top-right) LSF method, (bottom-left) conventional VKF and (bottom-right) proposed method.

For further interpretation of the proposed methodology, the envelope estimation is visualized in Figure 6 to assess the evaluation with respect to the reference one (actual envelope). It is essential to see how the model kept track of the envelope related to each of the 3 cyclic orders. In addition, the FRF along with the extracted component of the first cyclic order are displayed in Figures 7 and 8 to have a better comprehensive understanding of the model with respect to the chosen

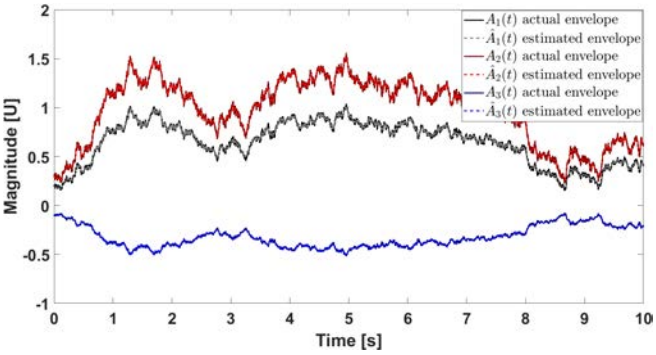


Figure 6. Plot showing the assessment of the estimated envelopes (dashed lines) with respect to the actual ones (full lines).

parameters β_k and ρ_1. It can be seen in the resiudal within the Figure 8 that the first tracked order was completely extracted after using the FRF displayed in 7. After iterating

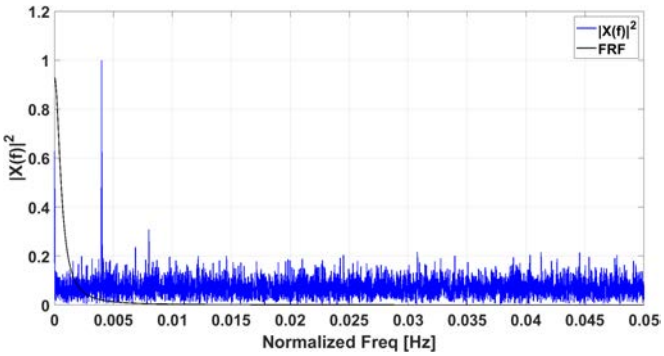


Figure 7. Scaled squared order spectrum of demodulated angular signal with FRF for first order, illustrating passed and rejected frequencies.

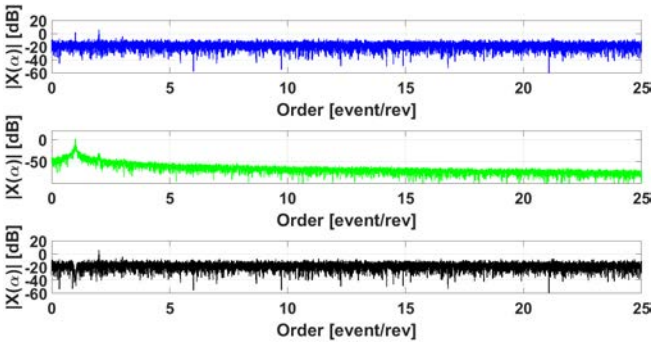


Figure 8. Order spectrums of: (top) the raw signal, (middle) the first extracted component and (bottom) the residual signal.

over the 3 orders, the spectrograms of the raw, extracted and residual signals are displayed in Figure 9. Since the signal is generated in a nonstationary regime, time-frequency representation (TFR) is a popular tool to present those time-variant components before resampling into the angular domain. This

is done to show that the proposed methodology filters well the 3 harmonics of interest from the raw signal.

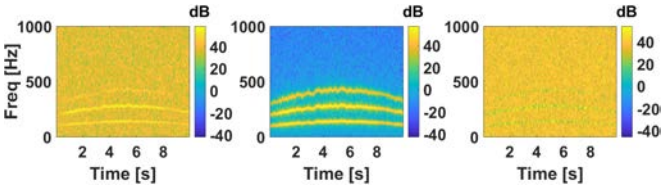


Figure 9. Simulation close-up spectrograms of the generated signals. (Left) Raw signal, (Middle) extracted components, and (Right) residual signal.

5. Application

This section presents the evaluation of the methodology using real experimental data acquired from the KU Leuven Diagnostic test rig (Yazdanianasr et al., 2024), as depicted in Figure 10. The test rig comprised an electric drive motor, a first housing containing a healthy bearing, and a second housing with two cases: one featuring large damage to the inner race of the test bearing and another case involving small damage to the outer race. The bearings used are SKF 2206 ETN9 bearings. The experimental setup also included the mounting of two ICP accelerometers (PCB-model number 352A10) on the housing of the bearings. Additionally, two B&k type 4188 microphones were installed as seen in the figure of the test rig. Furthermore, a smartphone, capturing through its microphone (considered as the low quality microphone in this acquisition), was placed behind the second microphone. Finally, an encoder was installed on the end of the electric motor to keep track of the angular position, providing a reliable estimate of the angular speed. Notably, the sampling frequency for all sensors, excluding the smartphone, was about 102.4 kHz. Given the smartphone's sampling frequency of 44.1 kHz, resampling was necessary to synchronize its data with that of the other sensors. The experiment aims to discern

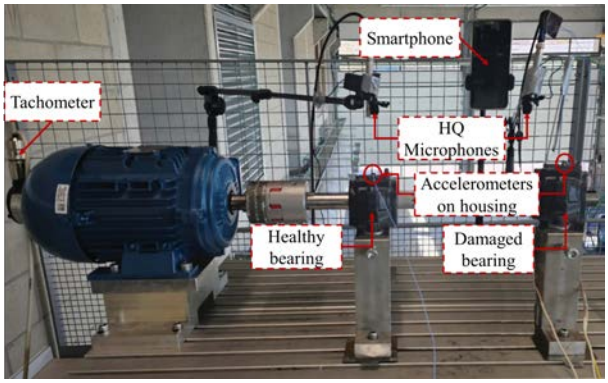


Figure 10. KU Leuven Diagnostic test rig.

both types of faults mentioned before, captured under a non-stationary regime. Naturally, signals from the accelerometer positioned on the upper part of the damaged bearing housing provide insights into significant damage to the inner race fault case. However, the experiment goes a step further by utilizing signals from the smartphone, which captures data from a distance near the upper part of the housing, revealing small damage for the outer race case. This small fault presents a significant challenge, as its detection amidst the presence of CS1 components can be particularly difficult to achieve, polluted by additional noise from the outside environment. Therefore, the experiment will be divided into two cases: the Large Inner Race Fault Case and the Small Outer Race Fault Case.

5.1. Large Inner Race Fault

In this initial experiment, the speed profile employed, as depicted in Figure 11, primarily exhibited random behavior to provide a highly nonstationary condition. The corresponding raw accelerometer signal is also displayed in Figure 12. In

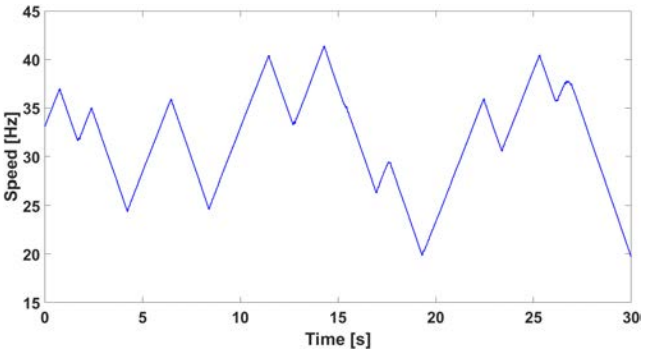


Figure 11. First experiment random walk-like speed profile

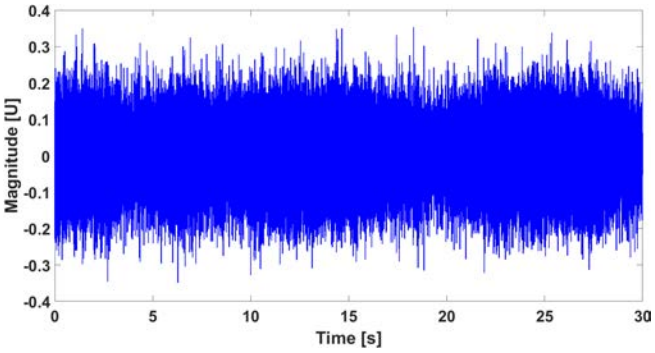


Figure 12. First experiment raw accelerometer signal depicting the varying signal envelope.

the initial phase, pinpointing the bandwidth containing critical CS2 information is essential. Utilizing the well known kurtogram (Antoni & Randall, 2007) can facilitate the detection of pertinent details. In this instance, the bandwidth yielding the highest kurtosis fell within the range of [815, 865] shaft order. Consequently, the CS1 orders to be tracked align with the speed components that reside within the frequency

band of [16300, 17300] Hz. After identifying 6 speed dependent orders, their deterministic components were extracted using the proposed technique. Upon fine-tuning the parameters of the methodology, the values shown in Table 3 were attained. The speed fluctuation weight λ was set the same for all harmonics to control the stationary-nonstationary adjustment. To evaluate the impact of the model with the specified parameters, close-up spectrograms of the raw signal, the extracted CS1 components, and the residual are presented in Figure 13. This enables observation of the significant attenuation of the speed components, tracked with the defined orders. Additionally, the speed profiles corresponding to the tracked orders are displayed with red dashed lines. The described processing steps are applied both to the raw signal and the residual one to compare the attenuation of speed dependent deterministic components. Initially, angular resampling is performed to mitigate frequency modulations. Subsequently, both angular signals undergo filtering within the specified bandwidth to isolate bearing signature information. A Hilbert transform is then employed to extract the envelope of the resulting complex signal. Finally, the squared envelope spectrum (SES) of the angular filtered signals is computed for evaluation. Figure 14 compares both SES, emphasizing the bearing fault contribution after the elimination of the extracted components. The analysis reveals that while extracting deterministic components, it also impacted the BPFI modulations. These modulations, typically associated with CS2 components, showed attenuation due to interactions with deterministic speed components. This suggests that the BPFI signature, not being entirely random, led to the attenuation of its deterministic part.

Table 3. Filter tuned parameters for the first experiment.

Parameters	Value
β_k	$\{0.995, 0.987, 0.999, 0.979, 0.988, 0.954\}$
ρ_k	$\{4320, 7664, 4211, 8853, 5333, 9912\}$
λ	0.01

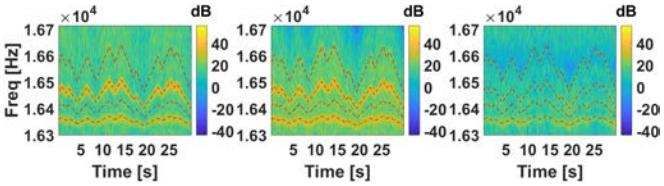


Figure 13. First experiment close-up spectrograms of the vibration signals. (Left) Raw signal, (Middle) extracted components, and (Right) residual signal with tracked speed profiles (red dashed lines).

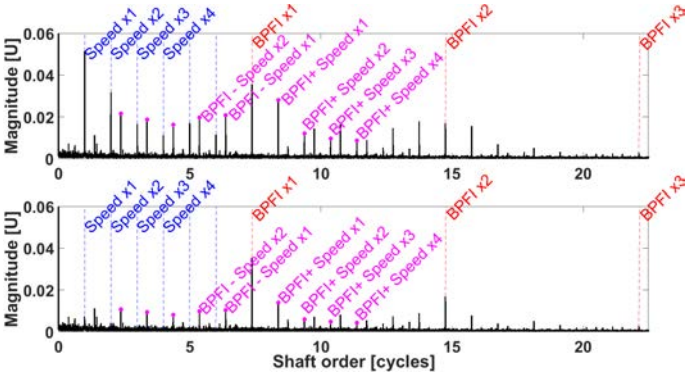


Figure 14. Squared envelope spectrum of: (top) raw signal and (bottom) residual signal highlighting the BPFI multiples (red dashed lines), the speed harmonics (blue dashed lines) and the modulations of the first BFPI (purple markers), showing the attenuation of the speed components in the residual one.

5.2. Small Outer Race Fault

In the second experiment, the utilized speed profile, illustrated in Figure 15, primarily demonstrated a steady-hop behavior. The corresponding raw smartphone microphone signal is also displayed in Figure 16 to demonstrate the variation of its envelope in tandem with the changes in speed. Similar to the initial steps taken in the first experiment, efforts were made to identify the bandwidth containing crucial CS2

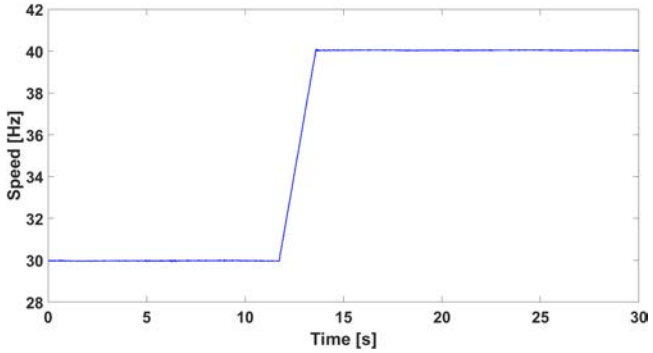


Figure 15. Second experiment steady-hop speed profile

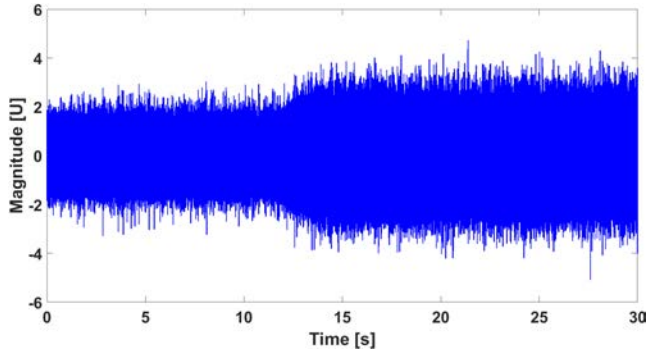


Figure 16. Second experiment raw smartphone microphone signal depicting the variation of the signal envelope mirroring the changes in the speed profile.

information. However, in this instance, the analysis was performed using data acquired from the smartphone instead of the accelerometer. The analysis revealed that the bandwidth with the highest kurtosis, as determined by the kurtogram, fell within the range of [333, 353] shaft order. For the second experiment, a similar process was followed to identify 8 speed dependent orders and extract their deterministic components using the proposed technique. The parameters of the methodology were fine-tuned to achieve the values described in Table 4. To assess the impact of the model with these specified parameters, close-up spectrograms of the raw signal, the extracted CS1 components, and the residual are provided in Figure 17 which facilitates observation of the significant attenuation of the speed components tracked with the defined orders.

Table 4. Filter tuned parameters for the second experiment.

Parameters	Value
β_k	$\{0.947, 0.955, 0.994, 0.936, 0.961, 0.887, 0.947, 0.984\}$
ρ_k	$\{12206, 11791, 8645, 13662 \; 9197, 21111, 12206, 10167\}$
λ	0.01

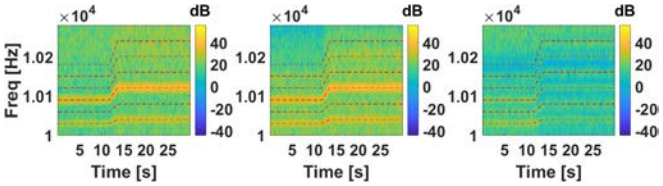


Figure 17. Second experiment close-up spectrograms of the vibration signals. (Left) Raw signal, (Middle) extracted components, and (Right) residual signal with tracked speed profiles (red dashed lines).

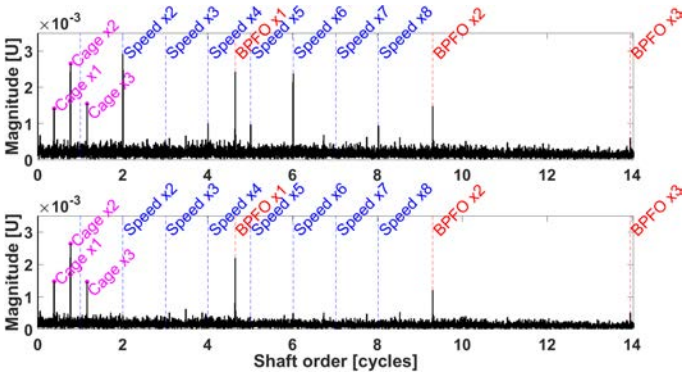


Figure 18. Squared envelope spectrum of: (top) raw signal and (bottom) residual signal highlighting the BPFO multiples (red dashed lines), the speed harmonics (blue dashed lines) and the cage frequencies (purple markers), showing the attenuation of the speed components in the residual one.

Similarly, for the second experiment, the processing until the computation of the SES was carried out on the raw signal and its residual to identify the Ball Pass Frequency Outer Race (BPFO). A comparison of both raw and residual SES are illustrated in Figure 18.

6. Conclusion

The paper introduced an extension of the Vold-Kalman Filter for better tracking of large nonstationary operating regimes. This extension is achieved by dynamically adapting the filter's bandwidth to accommodate fluctuations in speed. In the preliminary analysis, the frequency response function is also examined to provide insights into the filter's behavior. Numerical simulations are conducted to evaluate and compare the performance of conventional techniques against the proposed methodology. Additionally, the dependency on speed fluctuations is tested in a real-world application, specifically for enhanced bearing diagnostics. The adaptability to speed fluctuations ensures consistent and accurate performance across varying operational conditions, enhancing the effectiveness of machinery health monitoring. Looking ahead, future research could explore extending the proposed methodology to a higher filter order and automating the hyperparameter selection.

Acknowledgment

Fadi Karkafi gratefully acknowledges the European Commission for its support of the Marie Skłodowska Curie program through the ETN MOIRA project (GA 955681).

References

Abboud, D., Antoni, J., Sieg-Zieba, S., & Eltabach, M. (2016). Deterministic-random separation in nonstationary regime. *Journal of Sound and Vibration*, *362*, 305–326.

Abboud, D., Assoumane, A., Marnissi, Y., & El Badaoui, M. (2019). Synchronous fitting for deterministic signal extraction in non-stationary regimes: Application to helicopter vibrations.

Abboud, D., Marnissi, Y., Assoumane, A., Hawwari, Y., & Elbadaoui, M. (2022). Synchronous analysis of cyclo-non-stationary signals: A comprehensive study with aeronautic applications. *Mechanical Systems and Signal Processing*, *168*, 108600. doi: 10.1016/j.ymssp.2021.108600

Antoni, J. (2009). Cyclostationarity by examples. *Mech. Syst. Signal Process.*, *23*(4), 987–1036.

Antoni, J., & Randall, R. B. (2002). Differential diagnosis of gear and bearing faults. *J Vib Acoust*, *124*(2), 165–171.

Antoni, J., & Randall, R. B. (2007). Fast computation of the kurtogram for the detection of transient faults. *Mechanical Systems and Signal Processing*, *20*(1), 108–124.

Bonnardot, F., El Badaoui, M., Randall, R. B., Daniere, J., & Guillet, F. (2005). Use of the acceleration signal of a gearbox in order to perform angular resampling (with limited speed fluctuation). *Mech. Syst. Signal Process.*, *19*(4), 766–785.

Borghesani, P., Pennacchi, P., Randall, R. B., & Ricci, R. (2012). Order tracking for discrete-random separation in variable speed conditions. *Mech. Syst. Signal Process.*, *30*, 1–22.

Braun, S. (1975). The extraction of periodic waveforms by time domain averaging. *Acta Acustica United with Acustica*, *32*(2), 69–77.

Braun, S. (1986). Mechanical signature analysis: theory and applications.

Daher, Z., Sekko, E., Antoni, J., Capdessus, C., & Allam, L. (2010). Estimation of the synchronous average under varying rotating speed condition for vibration monitoring. *Journal of Sound and Vibration*.

Dion, J.-L., Stephan, C., Chevallier, G., & Festjens, H. (2013). Tracking and removing modulated sinusoidal components: A solution based on the kurtosis and the extended kalman filter. *Mechanical Systems and Signal Processing*, *38*(2), 428–439. doi: 10.1016/j.ymssp.2013.04.001

Feng, K., Ji, J., Wang, K., Wei, D., Zhou, C., & Ni, Q. (2022). A novel order spectrum-based vold-kalman filter bandwidth selection scheme for fault diagnosis of gearbox in offshore wind turbines. *Ocean Engineering*, *266*(Part 3), 112920. doi: 10.1016/j.oceaneng.2022.112920

McFadden, P. D. (1987). A revised model for the extraction of periodic waveforms by time domain averaging. *Mechanical Systems and Signal Processing*, *1*(1), 83–95.

McFadden, P. D. (1989). Interpolation techniques for time domain averaging of gear vibration. *Mech. Syst. Signal Process.*, *3*(1), 87–97.

Pai, P. F., & Palazotto, A. N. (2009, May 4-7). Online frequency and amplitude tracking of nonlinear non-stationary structural vibration. In *50th aiaa/asme/asce/ahs/asc structures, structural dynamics, and materials conference.* Palm Springs, California.

Pan, M., Chu, W., & Le, D.-D. (2016). Adaptive angular-velocity vold–kalman filter order tracking–theoretical basis, numerical implementation and parameter investigation. *Mech. Syst. Signal Process.*, *81*, 148–161.

Pan, M.-C., & Wu, C.-X. (2007a). Adaptive angular-displacement vold-kalman order tracking. In *2007 ieee international conference on acoustics, speech and signal processing - icassp '07* (pp. III-1293–III-1296). doi: 10.1109/ICASSP.2007.367081

Pan, M.-C., & Wu, C.-X. (2007b). Adaptive vold–kalman filtering order tracking. *Mech. Syst. Signal Process.*, *21*(8), 2957–2969.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—a tutorial. *Mech. Syst. Signal Process.*, *25*(2), 485–520.

Randall, R. B., Sawalhi, N., & Coats, M. (2011). A comparison of methods for separation of deterministic and random signals. *Int. J. Cond. Monit.*, *1*(1), 11–19.

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. doi: 10.1021/ac60214a047

Vold, H., Mains, M., & Blough, J. (1997). Theoretical foundations for high performance order tracking with the vold-kalman tracking filter. In *Sae technical paper.* Retrieved from `https://doi.org/10.4271/972007` doi: 10.4271/972007

Yazdanianasr, M., Verwimp, T., Karkafi, F., Mauricio, A., & Gryllias, K. (2024). *Acoustics dataset of damaged rolling element bearings captured using a smart phone at the ku leuven lmsd diagnostics test rig.* `https://doi.org/10.48804/XBN2QC`. KU Leuven RDR.

Enhancing Data-driven Vibration-based Machinery Fault Diagnosis Generalization Under Varied Conditions by Removing Domain-Specific Information Utilizing Sparse Representation

David Latil[1], Raymond Houé Ngouna[2], Kamal Medjaher[3], Stéphane Lhuisset[4]

[1,2,3] *Laboratoire Génie de Production, Université de Technologie Tarbes Occitanie Pyrénées, 47 Av d'Azereix, F-65016 Tarbes, France*
david.latil@uttop.fr
raymond.houe-ngouna@uttop.fr
kamal.medjaher@uttop.fr

[1,4] *Asystom, 244 Route de Seysses, Toulouse, France*
d.latil@asystom.com
s.lhuisset@asystom.com

ABSTRACT

This paper introduces a novel approach to machinery fault diagnosis, addressing the challenge of domain generalization in diverse industrial environments. Traditional methods often struggle with domain shift and the scarcity of balanced, labeled datasets, limiting their effectiveness across varied operational conditions. The proposed method leverages the abundance of healthy machinery signals as a reference for extracting domain-specific information. By doing so, it removes the domain-related variances from the observation signals, focusing on the intrinsic characteristics of faults. The methodology is validated with a case study, demonstrating enhanced diagnosis accuracy and generalization capabilities in unseen domains. This research contributes to the field of vibration-based intelligent fault diagnosis by providing a robust solution to a long-standing problem in machine condition monitoring.

1. INTRODUCTION

In the domain of industrial maintenance, ensuring the reliability and efficiency of rotating machinery is a central challenge. Among the various strategies employed, vibration-based fault diagnosis stands out as a proven technique for preemptive detection and mitigation of potential failures (Randall, 2010).

The advent of the Industrial Internet of Things (IIoT) and the proliferation of sensor technologies have led to an unprecedented availability of machinery data. This, in turn, has facilitated the application of intelligent diagnosis methods (Liu, Yang, Zio, & Chen, 2018) which showed impressive performance. Despite this, the use of these methods in real industrial scenarios has been proven difficult, mainly because it relies on a central assumption which is often hard to meet. Indeed, most Machine Learning (ML), including Deep Learning (DL) diagnosis techniques learn a representation of the training data in order to generalize to unseen examples. The unseen examples, also referred to as test data, must then follow the same distribution as the training data to ensure effective generalization by the model. The unpredictability of industrial environments and the varying working conditions of rotating machines significantly challenge this assumption. This results in overfitting on the working conditions the model has been trained on, and leads to a dramatic decrease in performance when conditions change.

Transfer learning has emerged as a popular strategy to address this challenge, aiming to leverage knowledge from one domain to improve performance in another. Specifically, methods employing distance metrics to bridge the gap between source and target domains have shown promise. However, these approaches typically assume availability of the target domain data during training, a scenario often impractical in the field. (Azari, Flammini, Santini, & Caporuscio, 2023). Indeed, despite the wide availability of surveillance data provided by IIoT sensors, the vast majority of available data predominantly reflects healthy working conditions, as faults are infrequent.

Domain generalization is then a more fitting problem formulation for situations where the target domain remains unknown during the model training phase. Unlike domain

adaptation, domain generalization aims to produce models which generalize well to domains unseen during training. For instance, in (Zhao & Shen, 2023) the authors proposed a mutual-assistance network for semi-supervised domain generalization, while in (Shi et al., 2023) a dynamic weighting strategy and a batch spectral penalization regularization term was employed to tackle the domain generalization problem. In (Jia, Li, Wang, Sun, & Deng, 2023) a deep causal factorization network is used, taking advantage of the causal properties in bearing signal models. The authors of (Zheng et al., 2021) combined apriori expert knowledge on vibration analysis and a deep neural network to generalize to unseen operating conditions. In (Wang et al., 2023) the authors used domain-specific discriminators to explicitly remove domain-specific information from the signals, creating domain-invariant representation, yielding to better generalization to unseen working conditions. However, the current landscape of domain generalization solutions is primarily characterized by complex deep learning architectures. Although effective, these architectures tend to obscure the interpretative transparency of these models, thus contributing to the 'black box' phenomenon often cited as a major pitfall of state-of-the-art models. Consequently, recent works such as (Kim et al., 2024) proposed an explainable diagnosis technique for single-domain generalization tasks using a priori knowledge to produce domain-invariant representations, showing increased performance on unseen target domains.

By tackling the domain shift challenge, our research introduces a novel preprocessing technique tailored to address the domain shift problem and the challenges induced by non-stationary vibration signals. This technique leverages the abundance of healthy signal data as a reference for identifying domain-specific information. We operate under the assumption that healthy signals contain such domain-specific information, which can impede the generalization capabilities of the models.

This approach aims to systematically eliminate domain-specific characteristics from the diagnosis data using advanced signal processing techniques, thereby isolating the intrinsic characteristics of faults. By focusing on the features that are truly indicative of machinery health, irrespective of operational conditions, our method proposes a step towards achieving domain-agnostic fault diagnosis. This approach allows us to benefit from the excellent performance state-of-the art intelligent models without increasing their complexity to achieve cross-domain fault diagnosis tasks.

The contributions of this paper are as follows:

1. A sparse representation-based signal processing technique is proposed to decompose the non-stationary noisy signals into their relevant components
2. Decomposed reference healthy signals are used to remove domain-specific information from the observation signals
3. The domain-invariant signals from a source domain are used train a simple classification model. Preprocessed signals from target domains are used to validate the generalization improvements on domains unseen during training.

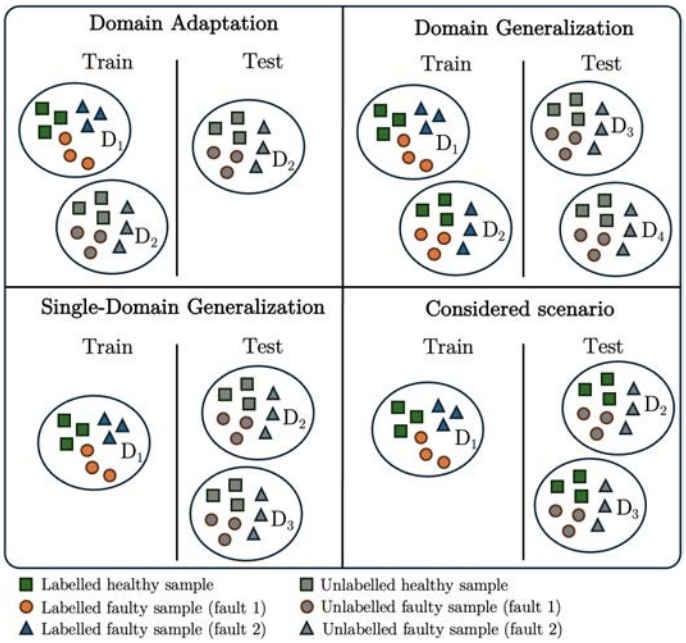


Figure 1. Generalization paradigms

The rest of this paper is organized as follows: in section 2 the method for domain-specific information removal is exposed, in section 3 an experimental setup and protocol is proposed to validate said method, and in section 4 the results are presented.

2. BACKGROUND

2.1. Domain generalization

Let us consider a rotating machine having N different working conditions, which translate into N different domains noted as $D^i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$, where (x_i, y_i) is the data-label pair for the jth sample in the ith domain. Let us also consider the situation where the label space is shared across domains, but only one domain is fully labelled and accessible during training, while in all others only healthy samples are known as such and are unseen until testing. This constitutes a realistic data availability scenario, where healthy data is abundant but fault data is scarce and usually represent a small subset of possible working conditions. The differences between domain adaptation, domain generalization, single-domain generalization and the Considered scenario are illustrated in Figure 1.

In this single-domain generalization scenario, the goal is to train a classifier considering the limited data availability and then demonstrate the generalization capability on unseen domains.

2.2. Vibration signals under time-varying working conditions

Rotating machines operating under varying working conditions not only cause the domain shift problem. Their vibration signals also contain specific challenges which makes them hard to process.

Vibration signals generated by rotating machines operating under constant or almost-constant working conditions can be described using Eq. 1.

$$x(t) = d(t) + r(t) + n(t), \tag{1}$$

where $d(t)$, $r(t)$ and $n(t)$ refer to deterministic, random and background noise contributions respectively. Under constant operating conditions, we can formulate several assumptions on the nature of these contributions. Deterministic contributions are almost-periodic as they are phase-locked to the shaft angle, and the random part is often described as cyclostationary (Antoni, Bonnardot, Raad, & El Badaoui, 2004), while background noise is often assumed to be Gaussian white noise coming from sensor and environmental noise.

Under varying operating conditions however, significant changes occur in the vibration signals of rotating machines which significantly challenge the assumptions previously made. For instance, when the rotating speed of the machine varies in time, the deterministic contributions are no longer periodic, while cyclostationary contributions become cyclo-non-stationary (Abboud et al., 2016). This emphasizes the enhancements outlined in this study, which will be discussed in the following section.

3. Proposed method

A preprocessing technique aiming to reduce the effects of domain shift induced by varying working conditions is proposed. The preprocessing technique is composed of two main tasks: the vibration signals must first be decomposed into their relevant parts, then the decompositions from reference signals are used to identify and remove domain-specific information from the signals in each domain. An overview schema of the method is illustrated in Figure 2.

3.1. Decomposition of vibration signals based on Sparse Representation

Many signal processing techniques have been proposed over the years to accurately handle vibration signals produced by rotating machines operating under time-varying working conditions. Considering the shortcomings of classical frequency-domain techniques, methods such as time-frequency analysis are often a suitable tool to handle these signals (Zhang & Feng, 2022). However each time-frequency has its own drawbacks, and choosing the right technique is often difficult. In this study, the Sparse Representation (SR) (Feng, Zhou, Zuo, Chu, & Chen, 2017) framework is adopted to decompose the vibration signals using a redundant basis.

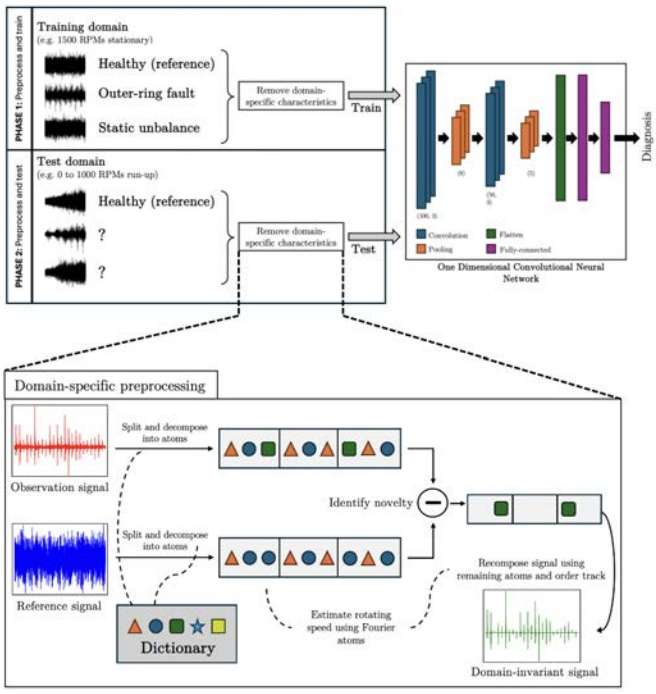


Figure 2. The proposed method to produce domain-irrelevant signals

Considering the morphological specificities of the different contributions present in vibration signals, SR allows not to be limited by the choice of a single basis, which might not be able to accurately represent all types of contributions. Instead a union of basis can be used, with the assumption that a more efficient and sparse representation can be achieved.

This union of basis is referred to as a dictionary, and each element of the dictionary is an atom. Several dictionaries have been proposed over the years, for instance the author of (Qin, 2018) used an impulse wavelet along with Fourier atoms to denoise bearing fault signals, while in (Cai, Selesnick, Wang, Dai, & Zhu, 2018) the authors used a union of a Discrete Cosine and a Short-time Fourier basis to diagnose faults in a gearbox.

These analytic dictionaries are very useful to identify components whose morphological characteristics, such as the natural frequency of the system, are known a priori. However in most cases there's a very limited amount of prior information

available on industrial machines. Therefore, this study adopts a minimalistic dictionary approach, accommodating both deterministic and stochastic elements in vibration signals. This is achieved through integrating Fourier and Unit bases, representing these contributions respectively.

SR can be achieved through either greedy methods like Matching Pursuit (MP) (Mallat & Zhang, 1993), or optimization-based techniques such as basis pursuit (BP) (Chen, Donoho, & Saunders, 1998). In the latter, the optimization objective is to minimize the reconstruction error, regularized by the norm of the sparse vector, expressed in Eq. 2.

$$\min_x \left\{ F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\psi(x) \right\}, \qquad (2)$$

where $y \in R^{N\times 1}$ is the input signal of size N, $A \in R^{N\times K}$ is the dictionary where $K > N$, x is the sparse vector, λ is the regularization parameter and ψ is a sparsity-inducing penalty.

In this study, the Generalized Minimax Concave (GMC) penalty is used due to its ability to overcome the amplitude underestimation issue associated with the l_1 penalty, while still preserving the convexity of the overall optimization objective, as highlighted by (Selesnick, 2017). The GMC penalty, defined in Eq. 3, serves as a key component in our approach.

$$\psi_{\text{GMC}}(x) = \|x\|_1 - \min_v \left\{ \|v\|_1 + \frac{\gamma}{2\lambda}\|A(x-v)\|_2^2 \right\}, \quad (3)$$

where $\gamma > 0$ controls the convexity of the GMC penalty, which is set at $\gamma = 0.8$ as advised in (Selesnick, 2017). The λ term is the regularization parameter. In this study, we set empirically $\lambda = 1.4$.

There are many algorithms designed to find the minimizer to this convex optimization problem. In this work we use the forward-backward splitting algorithm.

An example of decomposition using the proposed method is illustrated in Figure 3 where a signal containing a rolling element bearing fault is decomposed using Eq. 2. The different contributions from the Fourier and Unit basis are represented in blue and red respectively.

3.2. Removal of domain-specific components from the observation signals

After windowing and decomposing the signal using the proposed SR method, decompositions of healthy signals from each available domain are utilized as reference. It is assumed that these signals contain domain-specific characteristics that do not carry relevant diagnosis information and may contribute to the domain-shift issue. In every domain, atoms which are used to represent the reference signals are systematically removed from observation signals in order to produce domain-invariant signals, as illustrated in Figure 2.

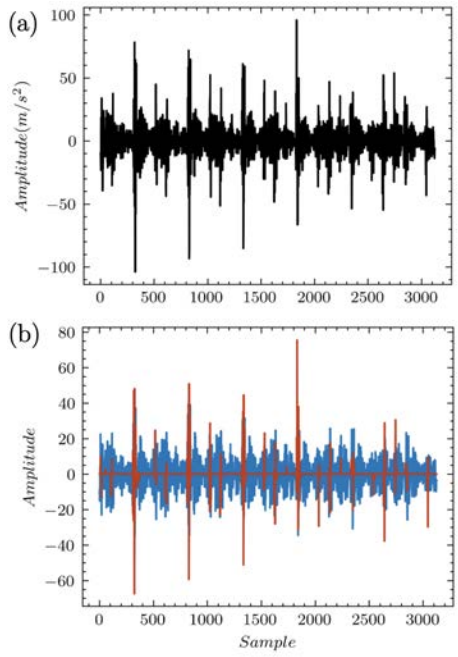


Figure 3. Decomposition of a signal containing an Inner Ring fault (a), into its deterministic contribution (blue in (b)) and random contributions (red in (b).

Additionally, the effects of varying speeds must be considered. Order Tracking is often used to resample the signal from the time domain to the order domain. However it requires information on the instantaneous rotating speed of the machine, which is often not available in industrial scenarios.

Consequently, we use the Fourier atoms from the sparse decompositions in order to estimate the instantaneous rotating speed without the need for additional hardware using the ridge tracking technique proposed in (Iatsenko, McClintock, & Stefanovska, 2016). The signal is then resampled from the time domain to the order domain, so that all domains share the same rotating speed reference.

4. Experimental verification

In this study, the Machinery Fault Simulator from SpectraQuest was used as test rig (pictured in Figure 3). It consists of an 3-phase 1HP motor, a main shaft with two Rexnord ER12K bearings and a gearbox linked to the main shaft by a double groove rubber belt. Three different couplings between the motor and the shaft are available (rigid, jaw, beam). A magnetic brake on the gearbox can be used to manually vary the load applied on the gearbox. The motor's speed can vary from 0 to 6000 RPMs.

Figure 4. The SpectraQuest test bench

Table 1. Considered faults

Defect	Type	Severity
Bearing	Inner ring	High
	Outer ring	High
Rotor	Static unbalance	Low
	Static unbalance	High

The vibration signals are acquired using three IFM VSA005 accelerometers sampling at 25.6 kHz, placed on the rightmost bearing housing. A high sampling rate is indispensable as some faults occur at high frequencies. Several artificial defects representative of most naturally occurring faults are introduced as summarized by Table 1.

In the present investigation, the test bench was employed to generate datasets across five distinct domains. Each domain is characterized by a specific speed curve that exemplifies an acceleration and deceleration cycle—commonly referred to as coast-up and coast-down phases. Such cycles are emblematic of the fluctuating operational conditions frequently encountered within industrial environments.

The domain shift problem is illustrated in Figure 5. A lightweight one-dimensional Convolutional Neural Network (1D-CNN) was used, based on the architecture described in Table 4, was trained on a single domain. The 1D-CNN is recognized as the state-of-the-art architecture (Borghesani, Herwig, Antoni, & Wang, 2023) for intelligent vibration-based fault diagnosis. Despite the impressive performance for working regimes of 1500 RPMs, the model accuracy drops significantly when the rotating speed varies.

Subsequently, five transfer tasks were defined, each depicted in Table III. The construction of these tasks allows defining actual transfer scenarios in the presence of varying working conditions.

The subsequent discussion will illustrate how the suggested pre-processing technique enhances the generalization capabilities of the CNN model, without requiring the adoption of

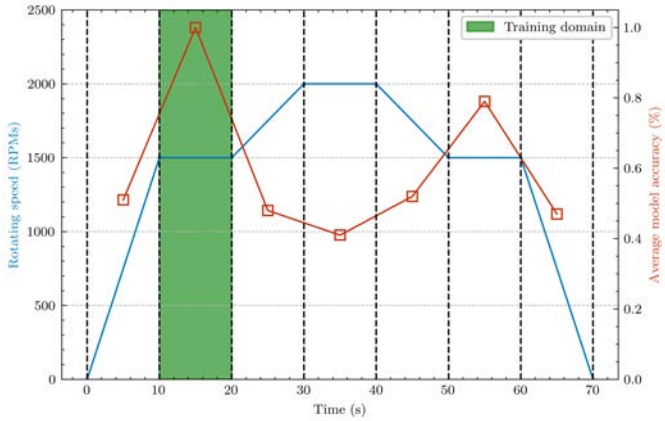


Figure 5. The domain shift problem: model accuracy decreases significantly when used on working conditions not represented in the training data

Table 2. Domains

Domain	Speed (RPM)
A	0 to 1500
B	1500
C	1500 to 2500
D	2500
E	2500 to 1500
F	1500
G	1500 to 0

Table 3. Cross-domain diagnosis tasks

Task	Source domain	Target domain
1	B	A
2	B	C
3	B	D
4	B	E
5	B	F
6	B	G

a more complicated model.

Table 4. 1D CNN Architecture

Layer Type	In. Ch.	Out. Ch.	Kernel/Stride/Size
Conv1d	1	3	Kernel=100, Stride=1
MaxPool1d	-	-	Kernel=8, Stride=8
Conv1d	3	3	Kernel=50, Stride=1
MaxPool1d	-	-	Kernel=5, Stride=5
Linear (FC1)	195	32	-
Dropout	-	-	p=0.5
Linear (FC3)	32	5	-

5. Results and Discussion

The model was trained in each task with 120 samples per source domain, with a sample being a 3125-long vibration signal in 10 different runs. The lambda parameter was set empirically to 1.2, the learning rate to 0.001, the number of epoch to 200. The early stopping strategy was employed to obtain a satisfying trained model. Note that the model itself is not the focus of the present study, it is merely used to demonstrate the generalization capabilities increase enabled by the proposed method.

The accuracies on unseen test domains with and without the preprocessing employed are then compared. It is important to note that whether with or without the preprocessing runs, the target domains were never included in the training data, meaning that the inference is performed on never-seen-before domain distributions.

The diagnosis results on each of the cross-domain diagnosis tasks are shown in Figure 4. Where it can be seen that the proposed pre-processing method increases the cross-domain accuracy of the model. The task 4 yields a diminished increase because the rotating speed of the target domain is identical to the source domain, showing that the proposed method does not decrease the adequate performance of in-distribution classification performance of modern models.

It must also be noted however that the first and sixth task's accuracy are not improved by the proposed method as very low decreasing speed carry very little energy and thus very little information, making it difficult to apply the proposed preprocessing scheme.

This study addressed the common issue of 'domain shift' caused by changes in a machine's operational environment that often lead to errors in machine fault detection by sophisticated computer models. The proposed approach sought to simplify the diagnosis process by filtering out the environmental noise from the signals machines give off, focusing in on the genuine indicators of malfunctions.

To validate the proposed technique, a test bench that simulates a variety of operational conditions and failures machines might encounter in real-world scenarios was utilized. Across

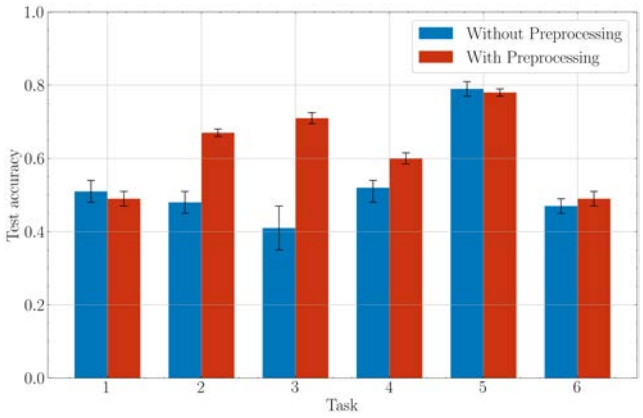


Figure 6. The effects of the proposed preprocessing method on the test accuracy of the model

five different domains, representing a range of typical industrial settings, our results indicate that our method, which employs a simple decision model with few parameters, was capable of identifying machine faults with an efficacy comparable to the more complex, state-of-the-art models currently in use.

6. Conclusion

In conclusion, this paper has presented a preprocessing technique that utilizes sparse representation to extract the domain-agnostic diagnosis information of machinery health signals, thereby significantly reducing the interference of domain-specific noise. The proposed method has been validated through a series of transfer tasks across different domains, revealing a significant improvement in model generalization without the necessity of resorting to complex neural network architectures.

On top of the generalization improvements, the proposed scheme use physically interpretable features which makes it easier to understand the output of the simple lightweight model employed here.

The study's limitations also open up new avenues for research, particularly in the domain of signal acquisition under extremely low-energy conditions. Addressing the shortfall in task 1 and 5 performance, where low decreasing speed results in signals with minimal information content, remains a challenge for future investigation.

References

Abboud, D., Baudin, S., Antoni, J., Rémond, D., Eltabach, M., & Sauvage, O. (2016, June). The spectral analysis of cyclo-non-stationary signals. *Mechanical Systems and Signal Processing*, *75*, 280–300. doi: 10.1016/j.ymssp.2015.09.034

Antoni, J., Bonnardot, F., Raad, A., & El Badaoui, M. (2004, November). Cyclostationary modelling of rotating machine vibration signals. *Mechanical Systems and Signal Processing*, *18*(6), 1285–1314. doi: 10.1016/S0888-3270(03)00088-8

Azari, M. S., Flammini, F., Santini, S., & Caporuscio, M. (2023). A Systematic Literature Review on Transfer Learning for Predictive Maintenance in Industry 4.0. *IEEE Access*, *11*, 12887–12910. doi: 10.1109/ACCESS.2023.3239784

Borghesani, P., Herwig, N., Antoni, J., & Wang, W. (2023, December). A Fourier-based explanation of 1D-CNNs for machine condition monitoring applications. *Mechanical Systems and Signal Processing*, *205*, 110865. doi: 10.1016/j.ymssp.2023.110865

Cai, G., Selesnick, I. W., Wang, S., Dai, W., & Zhu, Z. (2018, October). Sparsity-enhanced signal decomposition via generalized minimax-concave penalty for gearbox fault diagnosis. *Journal of Sound and Vibration*, *432*, 213–234. doi: 10.1016/j.jsv.2018.06.037

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998, January). Atomic Decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, *20*(1), 33–61. doi: 10.1137/S1064827596304010

Feng, Z., Zhou, Y., Zuo, M. J., Chu, F., & Chen, X. (2017, June). Atomic decomposition and sparse representation for complex signal analysis in machinery fault diagnosis: A review with examples. *Measurement*, *103*, 106–132. doi: 10.1016/j.measurement.2017.02.031

Iatsenko, D., McClintock, P., & Stefanovska, A. (2016, August). Extraction of instantaneous frequencies from ridges in time–frequency representations of signals. *Signal Processing*, *125*, 290–303. doi: 10.1016/j.sigpro.2016.01.024

Jia, S., Li, Y., Wang, X., Sun, D., & Deng, Z. (2023, June). Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *192*, 110228. doi: 10.1016/j.ymssp.2023.110228

Kim, I., Wook Kim, S., Kim, J., Huh, H., Jeong, I., Choi, T., … Lee, S. (2024, May). Single domain generalizable and physically interpretable bearing fault diagnosis for unseen working conditions. *Expert Systems with Applications*, *241*, 122455. doi: 10.1016/j.eswa.2023.122455

Liu, R., Yang, B., Zio, E., & Chen, X. (2018, August). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, *108*, 33–47. doi: 10.1016/j.ymssp.2018.02.016

Mallat, S., & Zhang, Z. (1993, December). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, *41*(12). doi: 10.1109/78.258082

Qin, Y. (2018, March). A New Family of Model-Based Impulsive Wavelets and Their Sparse Representation for Rolling Bearing Fault Diagnosis. *IEEE Transactions on Industrial Electronics*, *65*(3), 2716–2726. doi: 10.1109/TIE.2017.2736510

Randall, R. (2010, 12). Vibration-based condition monitoring: Industrial, aerospace and automotive applications. *Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications*. doi: 10.1002/9780470977668

Selesnick, I. (2017, September). Sparse Regularization via Convex Analysis. *IEEE Transactions on Signal Processing*, *65*(17), 4481–4494. doi: 10.1109/TSP.2017.2711501

Shi, Y., Deng, A., Deng, M., Li, J., Xu, M., Zhang, S., … Xu, S. (2023, June). Domain Transferability-Based Deep Domain Generalization Method Towards Actual Fault Diagnosis Scenarios. *IEEE Transactions on Industrial Informatics*, *19*(6), 7355–7366. doi: 10.1109/TII.2022.3210555

Wang, R., Huang, W., Lu, Y., Zhang, X., Wang, J., Ding, C., & Shen, C. (2023, October). A novel domain generalization network with multidomain specific auxiliary classifiers for machinery fault diagnosis under unseen working conditions. *Reliability Engineering & System Safety*, *238*, 109463. doi: 10.1016/j.ress.2023.109463

Zhang, D., & Feng, Z. (2022, January). Enhancement of time-frequency post-processing readability for nonstationary signal analysis of rotating machinery: Principle and validation. *Mechanical Systems and Signal Processing*, *163*, 108145. doi: 10.1016/j.ymssp.2021.108145

Zhao, C., & Shen, W. (2023, April). Mutual-assistance semisupervised domain generalization network for intelligent fault diagnosis under unseen working conditions. *Mechanical Systems and Signal Processing*, *189*, 110074. doi: 10.1016/j.ymssp.2022.110074

Zheng, H., Yang, Y., Yin, J., Li, Y., Wang, R., & Xu, M. (2021). Deep Domain Generalization Combining A Priori Diagnosis Knowledge Toward Cross-Domain Fault Diagnosis of Rolling Bearing. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11. doi: 10.1109/TIM.2020.3016068

Enhancing Gearbox Condition Monitoring using Randomized Singular Value Decomposition and K-Nearest Neighbor

Adel Afia[1,3], Moncef Soualhi[2], Fawzi Gougam[3], Walid Touzout [3], Abdassamad Ait-Chikh[4], and Mounir Meloussi[5]

[1] *Département de génie mécanique et productique, FGMGP, USTHB, 16111 Bab-Ezzouar, Algeria*
adel.afia@usthb.edu.dz
[2] *Université de Franche-Comté, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France*
moncef.soualhi@univ-fcomte.fr
[3] *LMSS, Faculté de technologie, Université de M'hamed Bougara Boumerdes, 35000 Boumerdes, Algeria*
f.gougam@univ-boumerdes.dz, w.touzout@univ-boumerdes.dz
[4] *LEMI, Faculté de technologie, Université de M'hamed Bougara Boumerdes, 35000 Boumerdes, Algeria*
ma.aitchikh@univ-boumerdes.dz
[5] *Faculté de technologie, Université de M'hamed Bougara Boumerdes, 35000 Boumerdes, Algeria*
m.meloussi@univ-boumerdes.dz

ABSTRACT

Efficient gear and bearing diagnosis has become a critical requirement across diverse industrial applications precisely due to their complex design and exposure to difficult operating conditions, which predispose them to frequent failure. Early fault identification remains problematic, as defects are commonly obscured by extensive background noise. Moreover, the exponential increases in gearbox data further complicate the defect classification process, confusing even the most sophisticated algorithms and significantly making the procedure time consuming. Singular Value Decomposition (SVD) has proved to be highly efficient in signal denoising, stability preservation, and feature extraction reliably under varying conditions, filtering out non-linear signals to reconstruct relevant features only. However, its considerable computation time necessitates exploring alternatives like Randomized SVD (RSVD) to mitigate processing time while maintaining classification accuracy. In this work, an intelligent algorithm for gear and bearing fault diagnosis is developed, incorporating Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) and Time-Domain Features for feature extraction. RSVD is employed for signal denoising and feature reconstruction, while K-Nearest Neighbor (KNN) for feature classification. The combined techniques ensure enhanced diagnostic accuracy, addressing critical challenges in industrial maintenance and performance optimization.

Keywords: Fault diagnosis, Gearbox, Feature extraction, Rotating machines.

1. INTRODUCTION

In rotating machines, particularly gearboxes, gears and bearings are susceptible to vulnerabilities due to their complex design and severe operating conditions which often compromise system reliability, leading to frequent failures requiring unscheduled maintenance and, ultimately, machine breakdowns. Notably in wind turbines, over 50% of gearbox faults come from bearings (de Azevedo et al., 2016), while approximately 80% of transmission machine problems are due to faulty gears (Soualhi et al., 2019). Consequently, the urgent need for machine fault diagnosis arises to ensure the safety and reliability of mechanical transmission systems. Moreover, today's competitive, dynamic and technology-driven industrial environment requires industry to adapt to new technologies, and to continually reduce costs (Benaggoune et al., 2020).

Intelligent fault diagnosis techniques primarily rely on machine monitoring parameters, with vibration analysis being a prevalent method for detecting early defects by identifying any deviations in these parameters. Vibration signals are especially favored due to their non-intrusive nature in machinery operation, making them a widely adopted tool for fault detection and analysis (Afia, Gougam, Rahmoune, et al., 2023). This approach enables continuous monitoring of machine health, allowing for timely interventions to prevent potential failures and to optimize maintenance schedules. Extracting fault-related characteristics from vibration signals poses a significant challenge, particularly in the initial fault development stages (Afia, Gougam, Rahmoune, et

al., 2023). Moreover, gears and bearings can incorporate a variety of defects, compounding the fault detection classification complexity. Recognizing these defects requires highlighting relevant information gleaned from measured vibration signals in a mathematically meaningful manner. Features serve as crucial signal characteristics aimed at encapsulating the overall data within a reduced dimensionality, facilitating their utilization in the classification process. Despite the complexities involved, effective feature extraction remains integral to accurately diagnosing faults and ensuring machinery reliability. Many decomposition methods have been developed for feature extraction. For instance, Gilles has proposed the empirical wavelet transform (EWT) (Afia, Gougam, Rahmoune, et al., 2023), in which input data is decomposed into multiple modes using a set of adaptive wavelet filters. The resulting EWT modes are narrow-band functions with fewer mixed modes, beneficial for many applications (Gilles, 2013). Nevertheless, EWT is highly dependent on the mode number selection, with improper selection potentially causing undesirable decomposition results (Adel et al., 2022). Furthermore, the wavelet filtering bandwidth adaptability in EWT is inherently limited, following a linear proportional bandwidth pattern (Adel et al., 2022). Discrete wavelet transform (DWT) is an alternative technique extensively used in fault diagnosis and condition monitoring (Syed & Muralidharan, 2022). DWT decomposes signal data through band pass filters in the time and frequency domains, producing a set of signals with specific frequency bands (Syed & Muralidharan, 2022). Yet, the dyadic step in the subsampling process represents a significant limitation in DWT efficiency (Adel et al., 2022). The Maximal overlap discrete wavelet transform (MODWT) has been developed as an optimized version of DWT to address the issue (Adel et al., 2022). Like DWT, MODWT invariably presents problems associated with poor frequency resolution [6]. As a solution, maximal overlap discrete wavelet packet transform (MODWPT) has appeared as a more suitable choice. MODWPT decomposes complex signals into individual components while maintaining circular shift equivariance, which is crucially important for gear and bearing condition monitoring (Adel et al., 2022). Moreover, MODWPT provides numerous improvements compared to MODWT, including uniform frequency bandwidths, the ability to overcome time-varying transformations, and to reconstruct the original signal without any information loss (Adel et al., 2022). MODWPT can extract relevant features from vibration data without compromising accuracy, thereby enhancing fault diagnosis processes.

Time-energy indicators such as kurtosis, entropy, root-mean square (RMS), etc., represent useful indicators in advanced signal processing algorithms for classifying different fault types (Soualhi et al., 2019; Gougam, Afia, Aitchikh, et al., 2024; Soualhi et al., 2020; Tahi et al., 2020). However, detecting bearing and gear signatures in early stages is extremely difficult as defects features are inherently weak. In such case, acquired vibration signals are often overwhelmed by a substantial amount of low-frequency noise, which makes significant impact on the analysis results' accuracy. For instance, in the event of local failure within the bearing, vibration signals exhibit a distinctly non-stationary behaviour, complicating the diagnostic process even more (Afia, Gougam, Touzout, et al., 2023). Consequently, achieving efficient fault identification continues to be an important issue in rotating equipment fault diagnosis. Addressing this issue is critical for enhancing the efficiency and accuracy of fault classification algorithms, necessitating strategies for noise reduction and optimization in feature selection processes. Singular value decomposition (SVD) is among the most commonly used methods, as highlighted in (Gougam et al., 2018; Touzout et al., 2020) due to its remarkable signal noise reduction and feature extraction capabilities, particularly in complex noise conditions. SVD is able to effectively reflect the matrix features since the singular values represent the intrinsic matrix features (Gougam et al., 2018; Touzout et al., 2020). Furthermore, SVD can maintain stability and improve the robustness of feature extraction under varying conditions. Since it is invariant, stable, and efficient for denoising, SVD has been used in practical applications, such as gear and bearing fault identification, to filter the nonlinear signal and ensure that only useful features are reconstructed. Despite its numerous advantages, the primary limitation of SVD lies in its high computational complexity. Addressing this challenge, Halko et al. proposed randomized SVD (RSVD) as an enhanced version of SVD (Halko et al., 2011). RSVD operates by generating an approximate basis for a range of input matrices through a process of "random sampling," wherein samples of the input matrix are multiplied by a random matrix (Song et al., 2017). This approach effectively captures the fundamental characteristics of the input matrix, including its singular values and most relevant vectors, reminiscent of data compression techniques. By enabling standard factorizations such as QR decomposition and SVD to be computed on a substantially smaller matrix than the original, RSVD significantly diminishes the computational cost (Song et al., 2017).

After feature extraction and noise reduction, K-Nearest Neighbor (KNN) has been widely adopted for gear and bearing fault detection and classification (Afia et al., 2024). The primary objective of this research is to investigate the effectiveness of a machine learning classifier in accurately classifying features extracted from vibration signals using MODWPT alongside temporal statistical indicators and RSVD. This paper presents a gear and bearing fault diagnosis method using vibration analysis, aiming to discern and categorize five distinct gear and bearing conditions. During the feature extraction phase, experimental vibration signals are decomposed by MODWPT, yielding several wavelet coefficients (WCs). Subsequently, 38 statistical features are applied to

each decomposed mode to construct a feature matrix corresponding to each gear and bearing condition. RSVD is then employed to reduce noise and to reconstruct feature matrix, ensuring the retention of pertinent features. Finally, KNN is utilized for feature classification, enabling the detection, classification, and identification of the five gear and bearing health states with precision and accuracy. This methodology represents a comprehensive approach towards enhancing gear and bearing fault diagnosis through advanced signal processing techniques and machine learning algorithm.

2. PROPOSED METHODOLOGY

In this section, the different steps of the proposed methodology are discussed. First, a total of 16 raw experimental vibration signals representing either a gear or a bearing state are decomposed using maximal overlap discrete wavelet packet transform (MODWPT) by 6 levels into 26 wavelet coefficients (WC) with different frequency levels. For one state, 16 matrices of (64×1048560) are produced, wherein 1048560 is the signal points number. Then, 38 combined time features are applied to each WC to construct the feature matrix corresponding to each condition. For one condition and one time feature, each matrix of (64×1048560) would be converted to a vector of 64 rows. Therefore, for one condition (16 measurements), a feature matrix (16×64) is provided to represent each gear or bearing condition. Combining 38 time features gives a feature matrix (608×64). After that, RSVD is used to reduce noise by calculating the right eigenvector, the singular value, and the left eigenvector in which Each feature matrix is reconstructed retaining the useful information only. The reconstructed feature matrices are used as inputs for KNN to detect, identify and classify the different states. To avoid over-fitting during the training and testing phases, 10-fold cross-validation is used, in which the dataset is randomly divided into 10 complementary subsets. Each subset is retained in turn, and the training model is trained on the remaining nine-tenths. Figure 1 provides an overview of the proposed technique.

3. MAXIMAL OVERLAP DISCRET WAVELET PACKET TRANSFORM

MODWPT uses raw data $X = [X_0, X_1, ..., X_{N-1}]^T$ as input for filtering and data decomposition. As with Mallat's algorithm (Gougam, Afia, Soualhi, et al., 2024; Too, Abdullah, Mohd Saad, & Tee, 2019), MODWPT is based on quadrature mirror filters. $\widetilde{g}_l$ and $\widetilde{h}_l q$ respectively represent a low-pass and a high-pass filters, each of length L (assumed to be even).Thus, the developed filters are given in Equation 1.

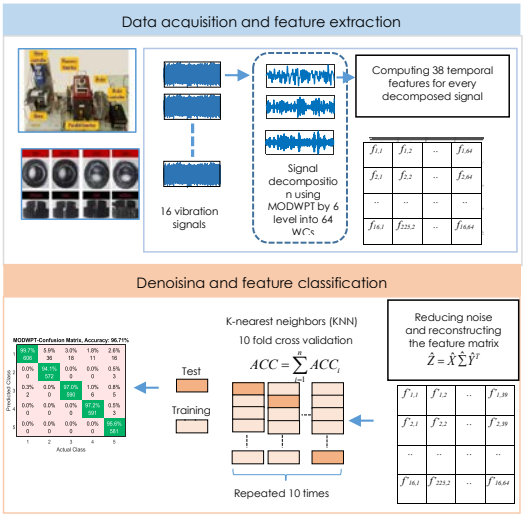
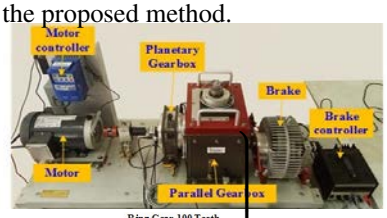
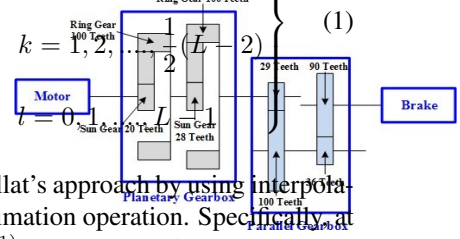


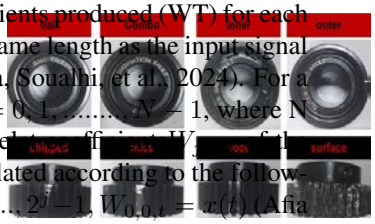
Figure 1. Diagram of the proposed method.

$$\begin{cases} \sum_{l=0}^{L-1} \tilde{h}_l^2 = \frac{1}{2} \\ \sum_{l=0}^{L-1-2k} \tilde{h}_l \tilde{h}_{l+2k} = 0, \quad k = 1, 2, \ldots, \frac{1}{2}(L-2) \\ \tilde{g}_l = (-1)^{l+1} \tilde{h}_{L-l-1}, \quad l = 0, 1, \ldots, L-1 \end{cases} \quad (1)$$





MODWPT differs from Mallat's approach by using interpolation instead of a 2-base decimation operation. Specifically, at each MODWPT level, $2^{(j-1)} - 1$ zeros are inserted between two consecutive adjacent coefficients of $\widetilde{g}_l$ *and* $\widetilde{h}_l$. Thereby ensuring that the wavelet coefficients produced (WT) for each wavelet sub-band maintain the same length as the input signal (Afia et al., 2024; Gougam, Afia, Soualhi, et al., 2024). For a discrete-time sequence $x(t), t = 0, 1, \ldots\ldots, N-1$, where N is the sequence length, the wavelet coefficient $W_{j,n}$ of the nth sub-band at level j are calculated according to the following equations in which $n = 0, 1..., 2^j - 1$, $W_{0,0,t} = x(t)$ (Afia et al., 2024; Gougam, Afia, Soualhi, et al., 2024):



$$\tilde{f}_{n,l} = \begin{cases} \tilde{g}_l, & \text{if } n \bmod 4 = 0 \text{ or } 3 \\ \tilde{h}_l, & \text{if } n \bmod 4 = 1 \text{ or } 2 \end{cases} \quad (2)$$

4. TEMPORAL FEATURES

The aim of this step of the methodology is to detect pattern changes in a given signal, in which statistical parameters are useful for extracting features related to the different machine states, since a failure will produce a change in the overall signal energy. For this purpose, 38 temporal features are used for feature extraction. The used time features are discussed in (Too, Abdullah, Mohd Saad, & Tee, 2019; Too, Abdullah,

& Saad, 2019).

5. RANDOMIZED SINGULAR VALUE DECOMPOSITION

Standard approaches use the extracted features from the previous step and directly train machine learning models for classification. However, achieving efficient fault classification accuracy seems to be a major issue in rotating equipment fault diagnosis, requiring noise reduction and optimization feature selection algorithms. In this situation, RSVD is used to reflect matrix features since singular values represent intrinsic matrix features, thus maintaining stability and improving the feature extraction reliability under varying conditions in practical applications, such as gear and bearing fault identification, by filtering the nonlinear signal and ensuring that only useful features are reconstructed with low computational complexity. For a matrix with m×n as dimension and k as rank, SVD gives this formula of$Z = XSY^*$, in which X is an orthonormal matrix $(m \times k)$, Y is an orthonormal matrix $(n \times k)$, while S is a non-negative diagonal matrix $(k \times k)$ which is defined in (Chakraborty et al., 2017):

$$W_{j,n,t} = \sum_{l=0}^{L-1} \tilde{f}_{n,l} \; W_{j-1,[n/2](t-2^{j-1}l) \bmod N} \tag{3}$$

$$S = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_k \end{bmatrix} \tag{4}$$

σ_j is the non-negative diagonal matrix S are the singular values of Z arranged as follows: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \sigma_4 \geq \sigma_k \geq 0$. The$XandY$ columns are the left and right singular vectors, respectively, while the singular values are related to the matrix approximation. At each level j, the number$\sigma_j + 1$ is equal to the spectral norm deviation between Z and an optimal rank-j approximation, in which (Too, Abdullah, & Saad, 2019):

$$\sigma_j + 1 = \min\{kZ - Bk \,:\, B\, has\, rank\, j\} \tag{5}$$

And the SVD of a matrix $Z \in R^{m \times n}$ is described as below (Chakraborty et al., 2017):

$$Z = X \sum Y^T \tag{6}$$

With X and Y being orthonormal, while $\sum$ is a rectangular diagonal matrix with diagonal entries being the singular values signified by σ_i. The column vectors of $XandY$representing the left and right singular vectors respectively, are indicated by$x_i and y_i$. In terms of $x_i and y_i$, the truncated SVD (TSVD) approximation of Z as a matrix Z_k is defined by (Chakraborty et al., 2017):

$$Z_k = \sum_{i=1}^{k} \sigma_i x_i {y_i}^T \tag{7}$$

And the randomized SVD (RSVD) is given as follow [23]:

$$\hat{Z} = \hat{X} \hat{\sum} \hat{Y}^T \tag{8}$$

In which and are each orthonormal while is diagonal that has as diagonal entries. The column vectors of and are referred as , and correspondingly. Elucidate the residual matrix of a TSVD approximation and the residual matrix of RSVD approximation are given below (Chakraborty et al., 2017):

$$R_k = Z - Z_k,\, and\, \hat{R}_k = Z - \hat{Z}_k \tag{9}$$

While the random projection of a matrix is elucidated as in (Too, Abdullah, Mohd Saad, & Tee, 2019):

$$Y = \Omega^T Z \;\; or \;\; Y = Z\Omega \tag{10}$$

In which Ω is a random matrix with independent and identically distributed entries. RSVD is an algorithm that examines approximate matrix factorization by employing random projections to divide the entire process into two steps. First, a random sampling is performed to obtain a reduced matrix with a range close to $Z's$ range. Thereafter, the reduced matrix is factorized using the first step on the matrix Z to find the orthonormal column matrix Q for $\xi > 0$ as defined in (Chakraborty et al., 2017):

$$\left\| Z - QQ^T Z \right\|_F^2 \leq \xi \tag{11}$$

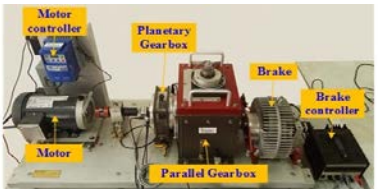
In the second step, the SVD of the reduced matrix $Q^T Z \in \mathbb{R}^{l \times m}$ is calculated, where $l \ll n$. Based on $\hat{X}\hat{\Sigma}\hat{Y}^T$ to denote the SVD of $Q^T Z$, Z is given in the following expression (Chakraborty et al., 2017):

$$Z \approx (Q\tilde{X}) \hat{\sum} \hat{Y}^T = \hat{X} \hat{\sum} \hat{Y}^T \tag{12}$$

Where $\hat{X} = Q\tilde{X}$ and $\hat{Y}$ are orthogonal matrices.

6. K-NEAREST NEIGHBORS

The reconstructed feature matrices are used as inputs for KNN to detect, identify and classify the different states. KNN is a simple and effective supervised classification approach, particularly in the field of pattern recognition, since it operates without the need for specific learning steps (Too, Abdullah, & Saad, 2019). When classifying a new input sample, KNN identifies the nearest neighbors of the training dataset and assigns the most common class to the new sam-



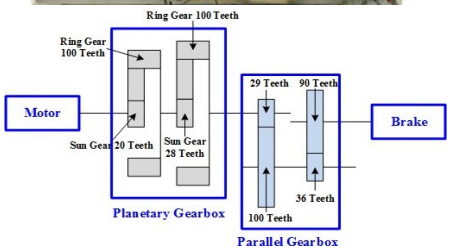


Figure 2. Gearbox setup schematic.

ple on the basis of a similarity measure. This process is conducted via unsupervised algorithmic methods, in which results are ranked on the basis of the majority of KNN categories (Anggoro & Kurnia, 2020). The algorithm works as follows:

1. **D**etermining the parameter k, representing the number of nearest neighbors.
2. **C**alculating the distance between the evaluated and the training data.
3. **S**orting the distances from high to low values.
4. **S**electing the nearest distances up to the order of k.
5. **A**ssigning the appropriate class based on the majority vote among the nearest neighbors.

7. Application and Results

The described methodology is applied to experimental data, which includes various fault states as well as a healthy state. The experimental setup is designed for multi- faults classification. With the proposed methodology, our objective is to evaluate the effectiveness of the extracted features in separating the different health states .

7.1. Case Study

components: motor, brake, planetary gearbox and parallel gearbox (see Figure 2) (Afia, Gougam, Rahmoune, et al., 2023). Defects (Table 1) were investigated in two distinct operating modes, with rotational speeds and loads ($20Hz - 0V$ and $30Hz - 2V$).

Eight 608A11 vibration sensors were placed on the test bench surface, with a 0.5 Hz-10 kHz frequency range, a $\pm 50g$ measurement range and 100 mV/g accuracy. Vibrations in the planetary gearbox directions were measured using three sensors, another three sensors to measure vibrations in the three

Figure 3. Gear and bearing defects.

Table 1. Types of bearing and gear components

Component	Types	Description
Gear	Chipped	Crack in the feet
	Miss	Missing
	Surface	Wear
	Root	Crack
Bearing	Ball	Crack
	Comb	Crack in inner and outer ring
	Inner	Crack
	Outer	Crack

directions of the parallel gearbox, and the remaining sensors monitored the drive motor. Load measurement was provided by an FT293 torque transducer with a measuring range of $\pm 5V$ and an accuracy of 4 Nm/V, placed between the motor and the planetary gearbox. Signal acquisition was achieved using a Spectra PAD compact data acquisition instrument able to process up to 20 channels, with a 1024 Hz sampling rate and a 512 second sampling window (Afia, Gougam, Rahmoune, et al., 2023).

7.2. Result and discussion

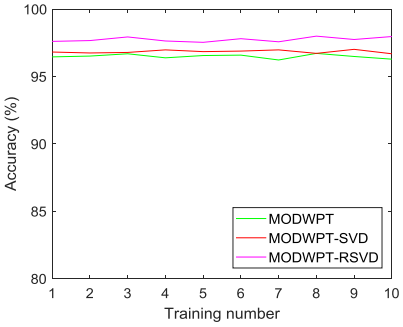
Raw vibration signals measured by the eight accelerometers corresponding to all five bearing and gear states for two operating modes (see TABLE I) are decomposed into 64 WCs using MODWPT. 38 time-based features are applied to each WC to create the feature matrices describing each gear or bearing's health state. Afterwards, RSVD computes right eigenvector, singular value and left eigenvector, and then each gear or bearing's feature matrix is reconstructed. The reconstructed matrices are taken as KNN inputs.

Model stability is an extremely important factor in determining potential model reliability in terms of overfitting, data variability or model sensitivity. By considering accuracy over repeated training iterations, a more complete model reliability assessment is provided. To evaluate the machine learning model's accuracy, TABLE II provides overall accuracy over ten training iterations using the proposed approach with and without RSVD. Figure 4 compares the model accuracy over ten training iterations, while Figs.5 and 6 provide a better illustration of the classifier's overall performance in terms of confusion matrices.

Compared with MODWPT and MODWPT-SVD, MODWPT-

Table 2. Classification accuracy of fa

Method	***Classification accuracy (%)***						
	Gear						
MODWPT	*96.45*	96.51	96.68	96.38	96.55	96.5	6.28
MODWPT-SVD	96.81	96.74	96.78	96.97	96.84	96.8	6.68
MODWPT-RSVD	97.60	97.66	97.93	97.63	97.53	97.8	7.96
	Bearing						
MODWPT	95.23	95.46	95.26	95.53	95.43	94.9	4.77
MODWPT-SVD	95.89	**96.12**	95.72	95.49	95.76	95.4	5.30
MODWPT-SVD	**97.27**	97.01	96.97	96.74	97.20	96.7	7.07



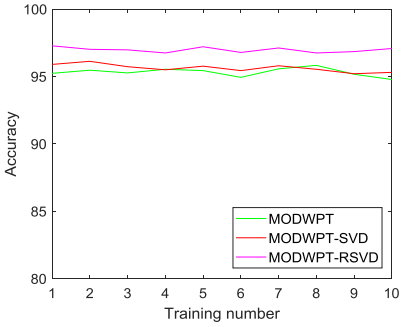


Figure 4. Model accuracy gear (a) bearing (b).

RSVD has achieved the best accuracy rates, mainly 97.99% for gears and 97.27% for bearings. This highlights our proposed method as a superior feature extraction technique, making it the optimal choice among the evaluated methods. Figure 4further confirms the proposed model's stability, providing highly satisfactory results in terms of fault classification. Thus, for accurate early defect detection and classification, MODWPT, with RSVD, provides the optimal approach.

8. Conclusion

The paper presents an enhanced fault diagnosis technique for gearboxes. Feature extraction, classification and experimental processes have been described in detail. The proposed algorithm, applied to real-tune gearbox vibration signals in different health states, successfully identified all gear and bear-

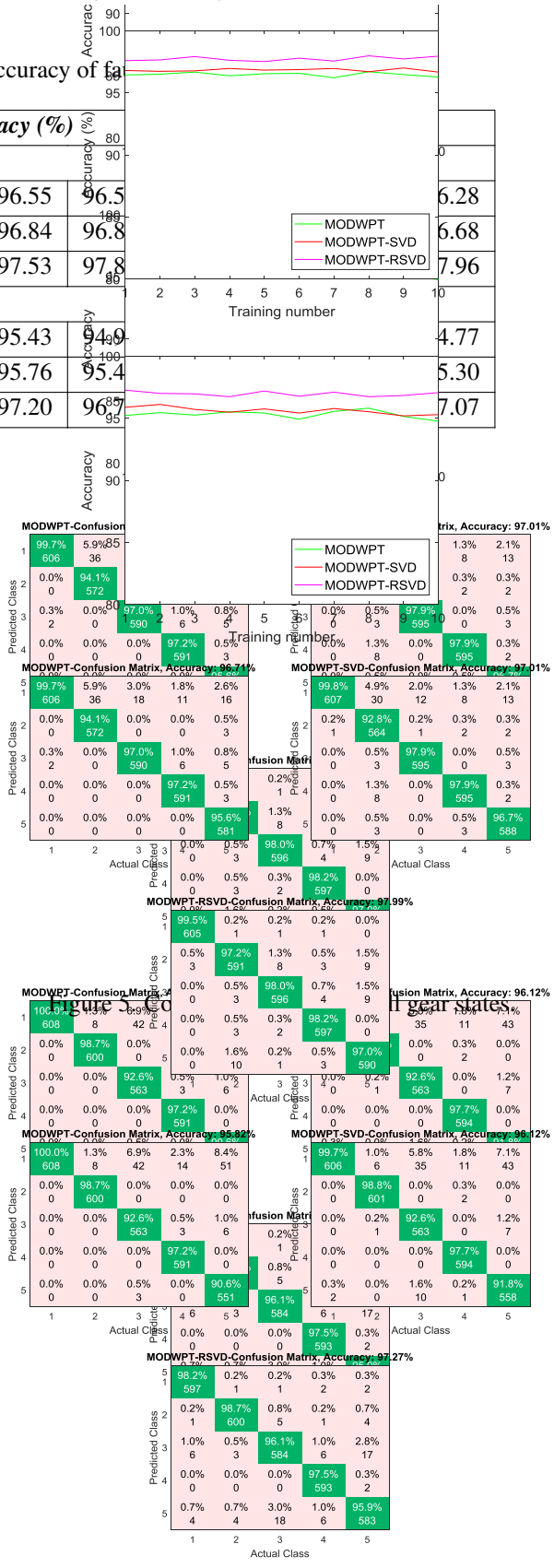


Figure 6. Confusion matrices for all bearing states.

ing states accurately and efficiently.

REFERENCES

Adel, A., Hand, O., Fawzi, G., Walid, T., Chemseddine, R., & Djamel, B. (2022). Gear fault detection, identification and classification using mlp neural network. In *Recent advances in structural health monitoring and engineering structures: Select proceedings of shm and es 2022* (pp. 221–234). Springer.

Afia, A., Gougam, F., Rahmoune, C., Touzout, W., Ouelmokhtar, H., & Benazzouz, D. (2023). Gearbox fault diagnosis using remd, eo and machine learning classifiers. *Journal of Vibration Engineering & Technologies*, 1–25.

Afia, A., Gougam, F., Rahmoune, C., Touzout, W., Ouelmokhtar, H., & Benazzouz, D. (2024). Intelligent fault classification of air compressors using harris hawks optimization and machine learning algorithms. *Transactions of the Institute of Measurement and Control*, *46*(2), 359–378.

Afia, A., Gougam, F., Touzout, W., Rahmoune, C., Ouelmokhtar, H., & Benazzouz, D. (2023). Spectral proper orthogonal decomposition and machine learning algorithms for bearing fault diagnosis. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *45*(10), 550.

Anggoro, D. A., & Kurnia, N. D. (2020). Comparison of accuracy level of support vector machine (svm) and k-nearest neighbors (knn) algorithms in predicting heart disease. *International Journal*, *8*(5), 1689–1694.

Benaggoune, K., Meraghni, S., Ma, J., Mouss, L., & Zerhouni, N. (2020). Post prognostic decision for predictive maintenance planning with remaining useful life uncertainty. In *2020 prognostics and health management conference (phm-besançon)* (pp. 194–199).

Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A. S., & Hassanien, A. E. (2017). Comparative approach between singular value decomposition and randomized singular value decomposition-based watermarking. *Intelligent techniques in signal processing for multimedia security*, 133–149.

de Azevedo, H. D. M., Araújo, A. M., & Bouchonneau, N. (2016). A review of wind turbine bearing condition monitoring: State of the art and challenges. *Renewable and Sustainable Energy Reviews*, *56*, 368–379.

Gilles, J. (2013). Empirical wavelet transform. *IEEE transactions on signal processing*, *61*(16), 3999–4010.

Gougam, F., Afia, A., Aitchikh, M., Touzout, W., Rahmoune, C., & Benazzouz, D. (2024). Computer numerical control machine tool wear monitoring through a data-driven approach. *Advances in Mechanical Engineering*, *16*(2), 16878132241229314.

Gougam, F., Afia, A., Soualhi, A., Touzout, W., Rahmoune, C., & Benazzouz, D. (2024). Bearing faults classification using a new approach of signal processing combined with machine learning algorithms. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *46*(2), 65.

Gougam, F., Rahmoune, C., Benazzouz, D., Zair, M. I., & Afia, A. (2018). Early bearing fault detection under different working conditions using singular value decomposition (svd) and adaptatif neuro fuzzy inference system (anfis). In *International conference on advanced mechanics and renewable energy (icamre). p* (pp. 28–29).

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, *53*(2), 217–288.

Song, P., Trzasko, J. D., Manduca, A., Qiang, B., Kadirvel, R., Kallmes, D. F., & Chen, S. (2017). Accelerated singular value-based ultrasound blood flow clutter filtering with randomized singular value decomposition and randomized spatial downsampling. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, *64*(4), 706–716.

Soualhi, M., Nguyen, K. T., & Medjaher, K. (2020). Pattern recognition method of fault diagnostics based on a new health indicator for smart manufacturing. *Mechanical Systems and Signal Processing*, *142*, 106680.

Soualhi, M., Nguyen, K. T., Soualhi, A., Medjaher, K., & Hemsas, K. E. (2019). Health monitoring of bearing and gear faults by using a new health indicator extracted from current signals. *Measurement*, *141*, 37–51.

Syed, S. H., & Muralidharan, V. (2022). Feature extraction using discrete wavelet transform for fault classification of planetary gearbox–a comparative study. *Applied Acoustics*, *188*, 108572.

Tahi, M., Miloudi, A., Dron, J., & Bouzouane, B. (2020). Decision tree and feature selection by using genetic wrapper for fault diagnosis of rotating machinery. *Australian Journal of Mechanical Engineering*.

Too, J., Abdullah, A. R., Mohd Saad, N., & Tee, W. (2019). Emg feature selection and classification using a pbest-guide binary particle swarm optimization. *Computation*, *7*(1), 12.

Too, J., Abdullah, A. R., & Saad, N. M. (2019). Classification of hand movements based on discrete wavelet transform and enhanced feature extraction. *International Journal of Advanced Computer Science and Applications*, *10*(6).

Touzout, W., Benazzouz, D., Gougam, F., Afia, A., & Rahmoune, C. (2020). Hybridization of time synchronous averaging, singular value decomposition, and adaptive neuro fuzzy inference system for multi-fault bearing diagnosis. *Advances in Mechanical Engineering*, *12*(12), 1687814020980569.

Enhancing Lithium-ion Battery Safety: Analysis and Detection of Internal Short Circuit basing on an Electrochemical-Thermal Modeling

Yiqi Jia[1], Lorenzo Brancato[1], Marco Giglio[1], Francesco Cadini[1,*]

[1] *Politecnico di Milano, Department of Mechanical Engineering, Via La Masa 1, 20156, Milan, Italy*
yiqi.jia@polimi.it
lorenzo.brancato@polimi.it
marco.giglio@polimi.it
**Corresponding author: francesco.cadini@polimi.it*

ABSTRACT

As the main cause of thermal runaway, the prompt identification of Internal Short Circuit (ISC) occurrences in lithium-ion batteries (LIBs) has emerged as a critical priority for ensuring battery safety. To address this critical need, for a comprehensive understanding of ISC behaviors, an electrochemical-thermal-ISC coupled model has been developed in this work to simulate battery performance across various ISC levels. This model is also utilized to validate the efficacy and robustness of the advanced detection approach proposed. By integrating both thermal and electrical aspects using the Pseudo Two-Dimensional (P2D) and Energy Balance Equation (EBE), our model serves as an efficient surrogate for ISC experiments. Key ISC indicators have been analyzed and integrated into the proposed ISC detection algorithm to enhance its effectiveness. The algorithm utilizes an Equivalent Circuit Model (ECM)-based approach for estimating ISC resistance. This research not only advances our understanding of ISC dynamics but also establishes a robust framework for the timely and reliable detection of ISCs. These advancements significantly enhance the overall safety and reliability of LIBs in electric vehicles (EVs).

1. INTRODUCTION

With the increasing growth and application of LIBs, particularly in EVs, concerns over battery safety have escalated due to a significant number of car fire accidents Chen et al. (2021). Among the recognized types of battery failure modes, ISC is considered the most significant safety concern for LIBs B. Liu et al. (2018).

While many studies have used mechanical abuse to induce ISC modes and quantify their effects on temperature, State of Charge (SoC), and other measurements, the precise mechanism of spontaneous ISC during the daily use of EVs remains unclear Huang et al. (2021). Therefore, early detection and accurate identification of ISC before it leads to thermal runaway (TR) have become key research areas.

According to Feng et al., generating failure data is a primary objective in developing a comprehensive online ISC detection approach Feng, Weng, Ouyang, and Sun (2016). Various methods have been employed in previous literature to induce ISC experimentally, including mechanical deformation like the nail penetration test Abaza et al. (2018) and crush tests Zhu, Zhang, Sahraei, and Wierzbicki (2016), as well as heating triggers such as inserting ISC devices within cells Orendorff, Roth, and Nagasubramanian (2011) and overheating Spinner et al. (2015). Additionally, dendrite growth and external short circuit (ESC) substitute tests have been used by L. Liu et al. (2020); Feng, He, Lu, and Ouyang (2018). Due to the challenges associated with reproducibility and safety in ISC experiments, researchers also opt to develop battery ISC models to capture ISC effects on main signals Kim, Smith, Ireland, and Pesaran (2012); Feng et al. (2016).

In this study, we generated ISC data by modeling a high-fidelity ISC model. Given that temperature growth and voltage drop are key ISC indicators Lai et al. (2021); Wu et al. (2023), we coupled a thermal and electrochemical model to simulate these responses for ISC detection algorithm development.

Another primary objective is the formulation of the detection algorithm. Achieving online and onboard diagnosis in EVs relies on signals measured by Battery Management Systems (BMS), necessitating a computationally efficient algorithm. Over recent years, several approaches leveraging voltage signals for ISC detection have been proposed, including observ-

ing abnormal voltage changes Keates, Otani, Nguyen, Matsumura, and Li (2010); Sazhin, Dufek, and Gering (2016); Seo, Goh, Park, Koo, and Kim (2017), capturing differences between predicted and actual values Yokotani (2014), and applying algorithms utilizing voltage signals Seo, Park, Song, and Kim (2020); Hu, Wei, and He (2020). Regarding another key ISC signal, temperature response, only limited works have utilized it through model-based approaches Feng, Ouyang, et al. (2018); Jia, Brancato, Giglio, and Cadini (2024).

In this study, we implemented the Extended Kalman Filter (EKF) algorithm based on a simplified lumped electrical-thermal model, as proposed in our previous work. Model parameters were estimated from a dataset generated by the high-fidelity plant model. Utilizing both the voltage and temperature signals, the direct indicator, R_{ISC}, was set as the state in the algorithm to be estimated to identify ISC levels.

The remainder of this paper elaborates on the detection approach in Section 2, provides a detailed description of the built ISC plant model in Section 3, presents the detection results and validation of the proposed approach in Section 4, and concludes in Section 5 by summarizing the findings and their implications.

2. An Overview of ISC Detection Approach

Figure 1 presents the comprehensive framework of the ISC detection approach proposed in this study.

As shown in the upper part of the figure, a coupled electrochemical thermal model (plant model) is developed to generate a dataset representing the operational behavior of a healthy battery, as detailed in subsequent sections. The Recursive Least Squares (RLS) parameter estimation tool is then employed to derive lumped model parameters, facilitating an accurate representation of battery electrical signals with computational efficiency for online detection algorithms.

The Equivalent Circuit Model (ECM) is chosen as the lumped electrical model due to its simplicity and widespread use in battery State of Charge (SOC) estimation in Battery Management Systems (BMS). The temperature lumped model, represented by Equation 1, incorporates heat generation from internal resistance and ISC resistance, as well as heat dissipation through natural air convection between the battery surface and the environment. To parameterize the ECM, the Hybrid Pulse Power Characterization (HPPC) working profile is applied to the plant model. The HPPC current and voltage simulated from the plant model are utilized for parameter estimation. Details of the estimated parameters applied for the detection algorithm are provided in Table 1.

$$mC_m \frac{dT}{dt} = \frac{V^2}{R_{ISC}} + R_0 I^2 - hA(T - T_a) \qquad (1)$$

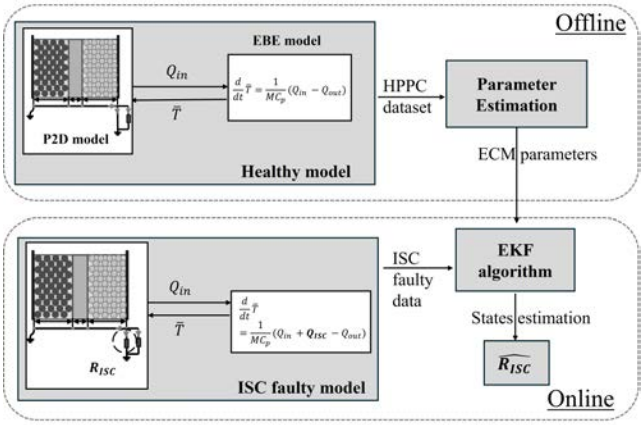


Figure 1. The overall ISC detection approach

Table 1. Battery parameters in lumped model

Symbol	Value	Unit
Q	2.3	Ah
R_0	0.0249	Ω
R_1	0.0072	**m**Ω
R_2	0.3641	**m**Ω
$\dot{\tau_1}$	13.18	s
$\dot{\tau_2}$	9.66	s

Figure 1 shows that, in the lower part (online detection phase), simulated ISC signals are obtained by introducing a parallel resistance as the ISC input for both the electrochemical and thermal models. The ECM-thermal-ISC model integrates parameters generated offline, including an additional item for ISC resistance, as a state to be estimated, into the model-based EKF algorithm for implementation and evaluation.

As highlighted by Hu et al. (2020), the state estimation process in the EKF consists of two primary stages: prediction and update. During the prediction stage, estimated state values are computed using model equations within the algorithm, incorporating the error covariance from the previous estimation step. This stage forecasts the next state based on the current state estimate and system dynamics. In the subsequent update stage, predicted states are refined by integrating measurements from sensors, which in this case are simulated values from the high-fidelity model.

The key aspect of the algorithm employed here involves incorporating the ISC resistance as one of the estimated states by integrating it into the ECM within the framework of the EKF algorithm. The state vector can be expressed as:

$$\mathbf{x} = [z,\ i_{R1},\ i_{R2},\ 1/R_{ISC}]^T, \qquad (2)$$

while the input and output vectors are:

$$\mathbf{u} = [i_t,\ v_t]^T, \qquad (3)$$

$$\mathbf{y} = [v_t,\ T]^T, \tag{4}$$

For further details regarding the functions and implementation of the algorithm, as well as other parameters applied, please refer to our previous work Jia et al. (2024).

3. Model Implementation

In this section, the detail of the electrochemical-thermal-ISC model developed to simulate the battery ISC is further described. The cell we simulated in this research is the A123 LiFePO4 26650.

3.1. Coupled Electrochemical-thermal Model

The electrochemical model employed is P2D model, which is based on a set of Partial Differential Equations (PDEs) describing the dynamics of physical processes within the battery electrodes and electrolyte Jokar, Rajabloo, Désilets, and Lacroix (2016). The main equations of the model are shown as below: Mass conservation of Li+ in the spherical active material:

$$\frac{\partial c_s}{\partial t} - \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 D_s \frac{\partial c_s}{\partial r}\right) = 0 \tag{5}$$

where c_s represents the concentration of Li+ in solid phase , r is the particle radius of the electrodes, D_s is the intercalation diffusivity.

Charge conservation in the electrodes:

$$\sigma_{\text{eff}} \frac{\partial^2 \phi_s}{\partial x^2} = j_f \tag{6}$$

where σ_{eff} is the effective elective electrical conductivity, ϕ_s is the electrical potential in solid phase, j_f is the electrode current density (I/A_s) and A_s is the specific interfacial area.

Mass conservation in the electrolyte phased:

$$\frac{\partial(\varepsilon_e c_e)}{\partial t} = \frac{\partial}{\partial x}\left(D_e^{\text{eff}} \frac{\partial c_e}{\partial x}\right) + \frac{1 - t_+^0}{F} j_f \tag{7}$$

where ε_e is the volume fraction of phase in electrolyte phased, c_e is the concentration of Li+ in electrolyte, D_e^{eff} is the elective electrolyte diffusivity, F is the Faraday's constant, t_+^0 is the transference number.

Charge conservation in electrolyte:

$$\frac{\partial}{\partial x}\left(\kappa^{\text{eff}} \frac{\partial \phi_e}{\partial x}\right) + \frac{\partial}{\partial x}\left(\kappa_{\text{D}}^{\text{eff}} \frac{\partial \ln c_e}{\partial x}\right) + j^{\text{Li}} = 0 \tag{8}$$

where κ^{eff} is the effective electrolyte conductivity, ϕ_e is the potential of the eletrolyte phase, j^{Li} is the reaction flux.

Over-potential and cell voltage:

$$\eta = \phi_s - \phi_e - U = \phi_s - \phi_e - (U_{\text{ref}} - (T - T_{\text{ref}})\frac{dU}{dT}) \tag{9}$$

where the T is the temperature of the battery cell.

The detail equations and the specific parameters used in this P2D model are sourced from the research by Prada et al. (2012). The temperature of this model is obtained from the output of the thermal model.

The thermal model is built based on the energy balance theory proposed by Bernardi et alBernardi, Pawlikowski, and Newman (1985). The temperature change with time can be described as Eq. 10

$$mC_m \frac{dT}{dt} = I(U - V - T\frac{dU}{dT}) - hA(T - T_{\text{a}}) \tag{10}$$

In this equation, the first part of the right side is the heat generation while it can be outputted from the electrochemical model, while I is the overall current, U is the open circuit voltage (OCV), V is the terminal voltage and $-T\frac{dU}{dT}$ is the reversible entropy change. The second part of the right is the heat dissipation while h, 10 $W/m^2/K$ is the heat transfer coefficient, A, 0.00634 m^2 is the inner surface area of the battery cell and T_{a}, 298 K is the environmental temperature. In the left side, m, 0.07 Kg is the mass of the battery cell and C_m, 1100 $J/kg/K$ is the heat capacity. These data are extracted from the battery data-sheet and other literature A123 Systems (2012); Song, Hu, Choe, and Garrick (2020); Bernardi et al. (1985).

The P2D electrochemical model and the thermal model can be coupled as shown in the figure 1, similar with Feng et al. (2016). As demonstrated in Eq.9 and Eq.10, the coupling achieved by considering the temperature dependent OCV. To be specific, the average temperature generated from the thermal model based on the heat generation and dissipation is timely converted to the electrochemical model by effecting the OCV.

3.2. ISC Model

As we illustrate in the figure 1, by paralleled the extra ISC resistance to the P2D, the ISC can be simulated. Therefore, the total current will be described as bellow:

$$I = I_{\text{t}} + I_{\text{ISC}} = I_{\text{t}} + \frac{V}{R_{\text{ISC}}} \tag{11}$$

At the same time, the heat generated from ISC counted from the Eq 12. should be added into the overall thermal equation.

$$Q_{\text{ISC}} = I^2 R_{\text{ISC}} \tag{12}$$

3.3. Simulation Result

By setting different values of the R_{ISC} value, various levels of ISC can be simulated based on the P2D-thermal-ISC model implemented here. Lower the R_{ISC} correspond to more se-

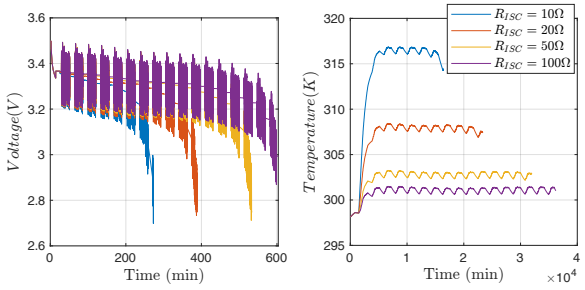


Figure 2. ISC simulation results

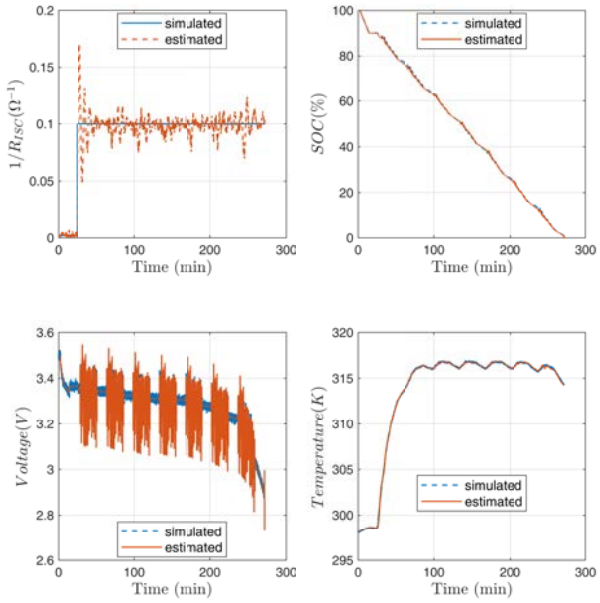


Figure 3. ISC detection results for $R_{ISC} = 10\Omega$

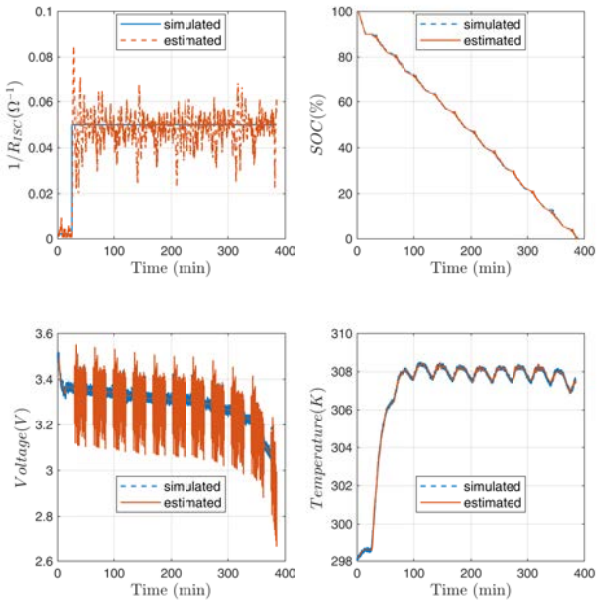


Figure 4. ISC detection results for $R_{ISC} = 20\Omega$

vere ISC conditions. The corresponding measurements simulated from the model are shown in figure 2 . Consistent with the findings in Feng, Ouyang, et al. (2018), the continual loss of SOC and the increase in heat generation are two main indicators of ISC occurrences. This is demonstrated by the quicker voltage drop and higher temperature rise observed with more severe ISC levels.

4. ISC Detection Results

To validate the proposed approach, measurements including voltage and temperature were generated from the built model using the Urban Dynamometer Driving Schedule (UDDS) profile. These signals were utilized by the algorithm to estimate the R_{ISC} online, with ISC simulated by specifying the R_{ISC} profile. The covariance of the measurement noises was set to be 0.1 mV and 2.5 mK, respectively.

For evaluating the performance of the early detection approach, we set R_{ISC} to a moderate level by dropping its value from a relatively high value (representing no ISC) to 10 Ω and 20 Ω, which are considered moderate ISC levels in other research studies (Feng, He, et al., 2018; Hu et al., 2020; Seo et al., 2020). The comparison between the algorithm's estimated states and simulated states from the built model is depicted in the top two sub-figures of Figure 3 and 4. Despite some fluctuations observed in the estimated R_{ISC}, these are attributed to measurement noise and the simplified model used within the EKF algorithm. Nonetheless, the estimated R_{ISC} closely and promptly tracks the simulated ISC value after its occurrence, demonstrating rapid and accurate detection of early ISC.

Specifically, from the results, it can be observed that after ISC is triggered, the estimated value converges to the true value within two minutes, while the temperature rises to 299 K for both cases. This implies that the approach can provide an alarm for severe ISC levels based on the estimated value of R_{ISC} before the temperature reaches a critical threshold, thereby preventing further thermal runaway or loss of battery capacity. Moreover, the SOC estimation displays good consistency with the real value. Furthermore, the voltage drop and temperature rise observed in the simulated results after ISC occurrence, as depicted in the figure, illustrate the primary ISC responses. These indicators demonstrate good consistency between the simulated data and the calculated data generated from the built-in model in the algorithm and the estimated states.

5. Conclusion

In this research, an electrochemical-thermal-ISC model was constructed in COMSOL to simulate ISC events and validate the proposed model-based ISC detection algorithm. This model combines the P2D model for electrical signals with the EBE for the thermal model, coupling them through temperature changes and the relationship between OCV and temperature.

The ECM-based algorithm demonstrated promising performance in early ISC detection. The ECM parameters utilized in the algorithm were derived through parameter estimation using a dataset generated from the high fidelity model under healthy battery conditions. .

In conclusion, the proposed method holds potential for application in BMS for early ISC detection, owing to its simplicity and efficiency. However, future research will focus on developing a more comprehensive battery model that considers the temperature response of physical electrochemical parameters to better capture dynamic responses and temperature distribution within the battery cell. Furthermore, enhancements to the algorithm will aim to incorporate temperature response and achieve precise localization of ISCs.

Acknowledgment

The author Yiqi Jia would like to thank the China Scholarship Council for the financial support (CSC, No.202108320086).

References

A123 Systems. (2012). *Nanophosphate high power lithium ion cell anr26650m1 b.* Data Sheet. (483 Specification, pp. 2–3)

Abaza, A., Ferrari, S., Wong, H. K., Lyness, C., Moore, A., Weaving, J., ... Bhagat, R. (2018). Experimental study of internal and external short circuits of commercial automotive pouch lithium-ion cells. *Journal of Energy Storage*, *16*, 211-217. doi: https://doi.org/10.1016/j.est.2018.01.015

Bernardi, D., Pawlikowski, E., & Newman, J. (1985). A general energy balance for battery systems. *Journal of the electrochemical society*, *132*(1), 5.

Chen, Y., Kang, Y., Zhao, Y., Wang, L., Liu, J., Li, Y., ... Li, B. (2021). A review of lithium-ion battery safety concerns: The issues, strategies, and testing standards. *Journal of Energy Chemistry*, *59*, 83–99. doi: 10.1016/j.jechem.2020.10.017

Feng, X., He, X., Lu, L., & Ouyang, M. (2018). Analysis on the fault features for internal short circuit detection using an electrochemical-thermal coupled model. *Journal of The Electrochemical Society*, *165*(2), A155.

Feng, X., Ouyang, M., Liu, X., Lu, L., Xia, Y., & He, X. (2018). Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. *Energy Storage Materials*, *10*(May 2017), 246–267. doi: 10.1016/j.ensm.2017.05.013

Feng, X., Weng, C., Ouyang, M., & Sun, J. (2016). Online internal short circuit detection for a large format lithium ion battery. *Applied Energy*, *161*, 168–180. doi: 10.1016/j.apenergy.2015.10.019

Hu, J., Wei, Z., & He, H. (2020). Improved internal short circuit detection method for Lithium-Ion battery with self-diagnosis characteristic. *IECON Proceedings (Industrial Electronics Conference)*, *2020-October*, 3741–3746. doi: 10.1109/IECON43393.2020.9254885

Huang, L., Liu, L., Lu, L., Feng, X., Han, X., Li, W., ... Ouyang, M. (2021). A review of the internal short circuit mechanism in lithium-ion batteries: Inducement, detection and prevention. *International Journal of Energy Research*, *45*(11), 15797–15831. doi: 10.1002/er.6920

Jia, Y., Brancato, L., Giglio, M., & Cadini, F. (2024). Temperature enhanced early detection of internal short circuits in lithium-ion batteries using an extended kalman filter. *Journal of Power Sources*, *591*, 233874. doi: https://doi.org/10.1016/j.jpowsour.2023.233874

Jokar, A., Rajabloo, B., Désilets, M., & Lacroix, M. (2016). Review of simplified pseudo-two-dimensional models of lithium-ion batteries. *Journal of Power Sources*, *327*, 44-55. doi: https://doi.org/10.1016/j.jpowsour.2016.07.036

Keates, A. W., Otani, N., Nguyen, D. J., Matsumura, N., & Li, P. T. (2010, September 14). *Short circuit detection for batteries.* Google Patents. (US Patent 7,795,843)

Kim, G.-H., Smith, K., Ireland, J., & Pesaran, A. (2012). Fail-safe design for large capacity lithium-ion battery systems. *Journal of Power Sources*, *210*, 243-253. doi: https://doi.org/10.1016/j.jpowsour.2012.03.015

Lai, X., Jin, C., Yi, W., Han, X., Feng, X., Zheng, Y., & Ouyang, M. (2021). Mechanism, modeling, detection, and prevention of the internal short circuit in lithium-ion batteries: Recent advances and perspectives. *Energy Storage Materials*, *35*(October 2020), 470–499. doi: 10.1016/j.ensm.2020.11.026

Liu, B., Jia, Y., Li, J., Yin, S., Yuan, C., Hu, Z., ... Xu, J. (2018). Safety issues caused by internal short circuits in lithium-ion batteries. *J. Mater. Chem. A*, *6*, 21475-21484. doi: 10.1039/C8TA08997C

Liu, L., Feng, X., Zhang, M., Lu, L., Han, X., He, X., & Ouyang, M. (2020). Comparative study on substitute triggering approaches for internal short circuit in lithium-ion batteries. *Applied energy*, *259*, 114143.

Orendorff, C. J., Roth, E. P., & Nagasubramanian, G. (2011). Experimental triggers for internal short circuits in lithium-ion cells. *Journal of Power Sources*, *196*(15), 6554–6558.

Prada, E., Di Domenico, D., Creff, Y., Bernard, J., Sauvant-

Moynot, V., & Huet, F. (2012). Simplified Electrochemical and Thermal Model of LiFePO 4 -Graphite Li-Ion Batteries for Fast Charge Applications . *Journal of The Electrochemical Society*, *159*(9), A1508–A1519. doi: 10.1149/2.064209jes

Sazhin, S. V., Dufek, E. J., & Gering, K. L. (2016, aug). Enhancing li-ion battery safety by early detection of nascent internal shorts. *ECS Transactions*, *73*(1), 161. doi: 10.1149/07301.0161ecst

Seo, M., Goh, T., Park, M., Koo, G., & Kim, S. W. (2017). Detection of internal short circuit in lithium ion battery using model-based switching model method. *Energies*, *10*(1), 76.

Seo, M., Park, M., Song, Y., & Kim, S. W. (2020). Online Detection of Soft Internal Short Circuit in Lithium-Ion Batteries at Various Standard Charging Ranges. *IEEE Access*, *8*, 70947–70959. doi: 10.1109/ACCESS.2020.2987363

Song, M., Hu, Y., Choe, S., & Garrick, T. (2020). Analysis of the heat generation rate of lithium ion battery using an electrochemical thermal model. *Journal of the Electrochemical Society*, *167*(12), 120503. doi: 10.1149/1945-7111/aba96b

Spinner, N. S., Field, C. R., Hammond, M. H., Williams, B. A., Myers, K. M., Lubrano, A. L., … Tuttle, S. G. (2015). Physical and chemical analysis of lithium-ion battery cell-to-cell failure events inside custom fire chamber. *Journal of Power Sources*, *279*, 713–721.

Wu, X., Wei, Z., Wen, T., Du, J., Sun, J., & Shtang, A. A. (2023). Research on short-circuit fault-diagnosis strategy of lithium-ion battery in an energy-storage system based on voltage cosine similarity. *Journal of Energy Storage*, *71*(May), 108012. doi: 10.1016/j.est.2023.108012

Yokotani, K. (2014). *Battery system and method for detecting internal short circuit in battery system.* Google Patents. (US Patent 8,643,332)

Zhu, J., Zhang, X., Sahraei, E., & Wierzbicki, T. (2016). Deformation and failure mechanisms of 18650 battery cells under axial compression [Article]. *Journal of Power Sources*, *336*, 332 – 340. (Cited by: 173) doi: 10.1016/j.jpowsour.2016.10.064

BIOGRAPHIES

Yiqi. JIA was born in China on January 28, 1996. She holds a Bachelor's degree in Automotive Engineering from Wuhan University of Technology, Wuhan, China (2017), and a Master's degree in Automotive Engineering from the University of Bath, Bath, UK (2018). After working as a vehicle engineer for nearly 3 years at Ford Motor Company, she started her Ph.D. journey in Mechanical Engineering at Politecnico di Milano in November 2021. Her research primarily focuses on the diagnosis and prognosis of Lithium-ion batteries, structural batteries, and more broadly, on mechanical/structural-related behaviors. This includes battery modeling, simulation, and data-based estimation methods for optimal battery management.

Lorenzo. BRANCATO was born in Italy on December 12, 1998. He earned his Bachelor's degree in Mechanical Engineering from Politecnico di Milano in 2020. Subsequently, he pursued a Master's degree in Mechatronic Engineering at Politecnico di Milano, completing his studies in 2022. He has started his Ph.D in Mechanical Engineering at Politecnico di Milano in 2023. His current research focus on the development of advanced diagnostic and prognostic approaches for dynamic, complex systems subject to degradation. This involves high-fidelity multi-physics modeling, simulation and advanced model-based filtering methods.

Marco. GIGLIO is Full Professor at the Department of Mechanical Engineering, Politecnico di Milano. His main research fields are: (i) Structural integrity evaluation of complex platforms through Structural Health Monitoring methodologies; (ii) Vulnerability assessment of ballistic impact damage on components and structures, in mechanical and aeronautical fields; (iii) Calibration of constitutive laws for metallic materials; (iv) Expected fatigue life and crack propagation behavior on aircraft structures and components; (v) Fatigue design with defects. He has been the coordinator of several European projects: HECTOR, Helicopter Fuselage Crack Monitoring and prognosis through on-board sensor, 2009-2011, ASTYANAX (Aircraft fuselage crack Monitoring System and Prognosis through eXpert on-board sensor networks), 2012-2015, and SAMAS (SHM application to Remotely Piloted Aircraft Systems), 2018-2020. He has been the project leader of the Italian Ministry of Defence project in the National Plan of Military Research, SUMO (Development of a predictive model for the ballistic impact), 2011-2012 and, SUMO 2 (Development of an analytical, numerical and experimental methodology for design of ballistic multilayer protections), 2017-2019. He has published more than 210 papers, h-index 27 (source Scopus) in referred international journals and congresses.

Francesco. CADINI (MSc in Nuclear Engineering, Politecnico di Milano, 2000; MSc in Aerospace Engineering, UCLA, 2003; PhD in Nuclear Engineering, Politecnico di Milano, 2006) is Associate Professor at the Department of Mechanical Engineering, Politecnico di Milano. He has more than 20 years of experience in the assess-

ment of the safety and integrity of complex engineering systems, entailing (i) artificial intelligence (machine learning)-based approaches for classification and regression, (ii) development and application of advanced Monte Carlo algorithms for reliability analysis (failure probability estimation), (iii) diagnosis and prognosis (HUMS) of dynamic, complex systems subject to degradation, (iv) uncertainty and sensitivity analyses, (v) structural reliability analyses.

Enhancing Lithium-Ion Battery State-of-Charge Estimation Across Battery Types via Unsupervised Domain Adaptation

Mohammad Badfar[1], Ratna Babu Chinnam[1], and Murat Yildirim[1]

[1] *Department of Industrial and Systems Engineering, Wayne State University, Detroit, Michigan, 48201, United States*
mohammadbadfar@wayne.edu
ratna.chinnam@wayne.edu
murat@wayne.edu

ABSTRACT

Accurate estimation of the state-of-charge (SOC) in lithium-ion batteries (LIBs) is paramount for the safe operation of battery management systems. Despite the effectiveness of existing SOC estimation methods, their generalization across different battery chemistries and operating conditions remains challenging. Current data-driven approaches necessitate extensive data collection for each battery chemistry and operating condition, leading to a costly and time-consuming process. Hence, there is a critical need to enhance the generalization and adaptability of SOC estimators. In this paper, we propose a novel SOC estimation method based on Regression-based Unsupervised Domain Adaptation. We evaluate the performance of this method in cross-battery and cross-temperature SOC estimation scenarios. Additionally, we conduct a comparative analysis with a widely-used classification-based unsupervised domain adaptation approach. Our findings demonstrate the superiority of the regression-based unsupervised domain adaptation method in achieving accurate SOC estimation for batteries.

1. INTRODUCTION

Accurate real-time estimation of the state-of-charge (SOC) in batteries holds paramount importance across various domains, including electric vehicles and renewable energy storage systems. The SOC represents the percentage of remaining capacity, serving as a pivotal indicator of the battery's condition for facilitating effective operations.

Precise SOC estimation is imperative for optimizing energy utilization and mitigating premature degradation, consequently reducing maintenance costs and environmental impacts. However, SOC determination poses a formidable challenge due to its dependence on multiple interconnected variables such as voltage, current, resistance, and temperature, complicating precise estimation (Z. Wang, Feng, Zhen, Gu, & Ball, 2021). Thus, the development of robust and adaptable SOC estimation methods is essential to meet the escalating demand for sustainable energy solutions. Conventional SOC estimation approaches often falter in dynamic environments characterized by temperature variations, load fluctuations, and battery aging. Compounding this challenge is the diverse array of battery types with varying chemistries. Conventional methods necessitate significant investments in time and resources to acquire labeled data specific to each battery variant for accurate SOC estimation. Consequently, there arises a crucial need for innovative, adaptable SOC estimation methods capable of addressing these challenges while reducing reliance on expensive labeled data sources.

Numerous methodologies have been proposed for SOC estimation, employing diverse sensor data and modeling techniques. Traditional approaches, such as look-up table methods and direct-counting methods, often rely on simple algorithms but struggle with real-time estimation due to their requirement for stable discharge currents (Shen, Li, Meng, Zhu, & Shen, 2023). Conversely, model-based methods address this limitation but demand prior knowledge of battery characteristics, rendering them less suitable for dynamic and varied operational conditions. Recent advancements have introduced data-driven methods, which eschew reliance on domain knowledge and instead utilize battery parameters such as current, voltage, and temperature measurements to develop SOC estimators. Various data-driven techniques have been proposed for battery SOC estimation. (Li, Wang, & Gong, 2016; Hu et al., 2014; Tong, Lacap, & Park, 2016; Khumprom & Yodo, 2019; Chandra Shekar & Anwar, 2019). How et al. offer a comprehensive review of SOC estimation methods (How, Hannan, Lipu, & Ker, 2019). The primary drawback of data-driven approaches lies in their dependence on substantial training data, which can be expensive and time-intensive to acquire. In response to the challenge of limited data, transfer learning (TL) has emerged as a potent technique in machine

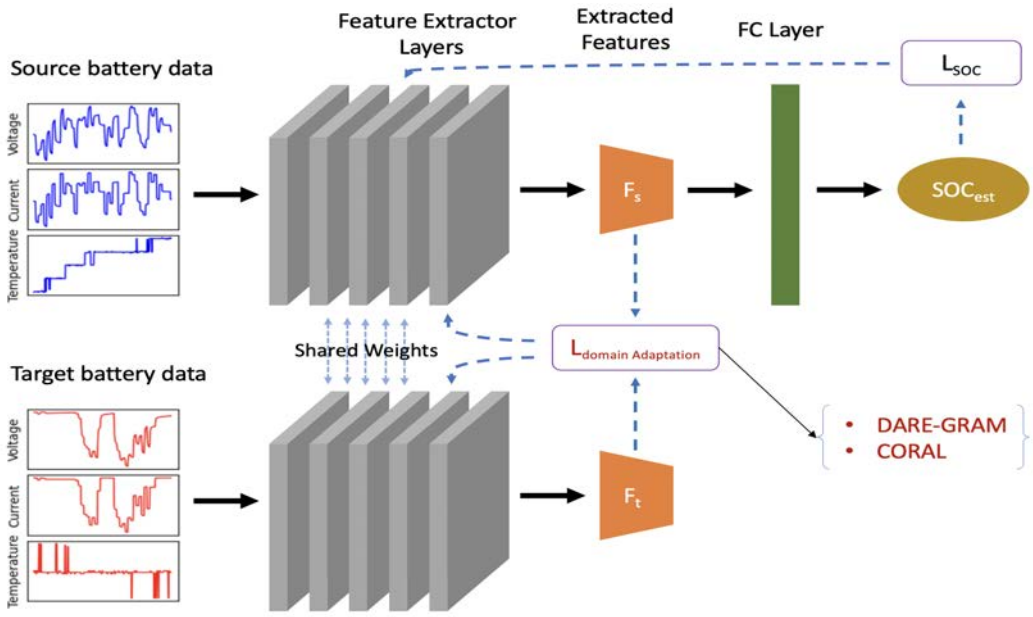


Figure 1. Structure of the Proposed Deep Neural Network Architecture for SOC Estimation using Domain Adaptation.

learning.

In the realm of battery SOC estimation, transfer learning holds promise by leveraging existing data from one domain (e.g., a specific battery type or environment), known as the source domain, to enhance SOC estimation performance in a different, less well-characterized domain, referred to as the target domain. Fine-tuning, the most popular TL approach, involves further training a pre-trained neural network model, originally trained on a large dataset for a different domain, using a smaller dataset specific to the target domain. Fine-tuning has recently been applied to battery SOC estimation to transfer knowledge between different ambient temperatures of the same battery type (Y.-X. Wang, Chen, & Zhang, 2022), or between different battery types (Bhattacharjee, Verma, Mishra, & Saha, 2021). However, fine-tuning necessitates access to labeled examples from the target domain, which may not always be readily available. In the case of lithium-ion batteries (LIBs), obtaining reliable labeled data under real-world conditions is particularly challenging.

To address the challenge of lacking labeled data for the target domain, machine learning researchers have introduced Unsupervised Domain Adaptation (UDA). Originating in the computer vision domain, UDA tackles the broader issue of transferring knowledge from a source domain to a target domain where labeled data is scarce (Long, Cao, Wang, & Jordan, 2015). This is particularly pertinent in scenarios where the characteristics of the target domain evolve over time, diverging from the source domain, and making traditional supervised learning approaches inadequate. Batteries are subjected to diverse environmental conditions, undergo degradation over time, and witness frequent introductions of new battery chemistries. A central strategy of UDA techniques is to generate domain-invariant feature representations by aligning feature distributions between domains (Wilson & Cook, 2020). This facilitates model adaptation to new and dynamically changing environments, enabling effective generalization without access to labeled data in the target domain.

A common approach for generating domain-invariant feature representations is to minimize a divergence measured as the distance between distributions. Maximum mean discrepancy (MMD) (Borgwardt et al., 2006), multi-kernel MMD (MK-MMD) (Gretton et al., 2012), and lastly, correlation alignment (CORAL) (Sun & Saenko, 2016) are among the popular divergence minimization techniques. Recently, there has been growing interest in applying UDA techniques to estimate battery SOC (Shen, Li, Liu, Zhu, & Shen, 2022; Bian, Yang, & Miao, 2020; Oyewole, Chehade, & Kim, 2022; Ni, Li, & Yang, 2023; Meng, Agyeman, & Wang, 2023). While these UDA techniques were initially developed for classification tasks, SOC estimation poses a regression task. A significant distinction between regression and classification problems is that regression problems are less robust to feature scaling, potentially impacting model robustness when aligning feature distributions with UDA methods (Chen, Wang, Wang, & Long, 2021).

To address this challenge, a specialized domain adaptation method for regression problems has emerged. DARE-GRAM (Nejjar, Wang, & Fink, 2023) is one such recent domain adaptation regression (DAR) technique motivated by the closed-form solution of ordinary least squares (OLS). Unlike pre-

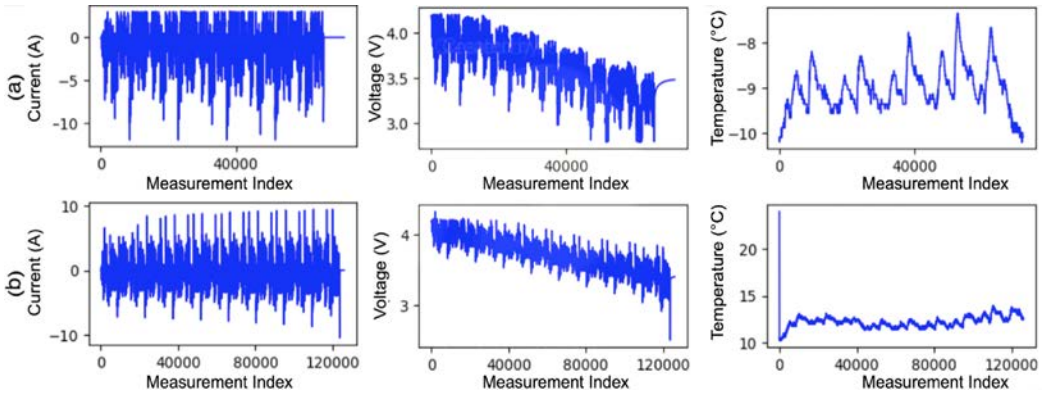


Figure 2. Measured Current, Voltage, and Battery Temperature During LA92 Drive Cycle. (a) Panasonic LiB and (b) LG LiB.

viously discussed classification-based methods that directly align features, DARE-GRAM aligns the inverse Gram matrix of the features. The authors demonstrated the capability and robustness of this method through experiments on three benchmark computer vision regression datasets.

In this paper, we explore the application of unsupervised domain adaptation (UDA) techniques within the framework of transfer learning (TL) to enhance the precision of battery State of Charge (SOC) estimation. We conduct a comparative study between a well-established classification-based domain adaptation method (CORAL) and DARE-GRAM, a regression-based method, marking the first application of a regression-based UDA method to battery management tasks, to the best of our knowledge. We examine their efficacy across a range of TL tasks and settings, aiming to provide comprehensive insights into their performance and suitability for addressing the complex challenges posed by evolving battery landscapes and diverse operational conditions.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology. Section 3 elucidates the LiB datasets and implementation details. Section 4 offers the experimental results and discussion. Finally, Section 5 concludes the paper.

2. METHODOLOGY

2.1. Problem Statement

We begin by defining the problem of cross-battery state-of-charge (SOC) estimation. The source domain D^s represents the battery type with labeled data $X^s = \{x_j^s, y_j^s\}_{j=1}^{n_s}$, where n_s denotes the number of source samples. Conversely, the target domain D^t represents the battery type with unlabeled data $X^t = \{x_j^t\}_{j=1}^{n_t}$, where n_t represents the number of target samples. x_j^s and x_j^t are the temporal measurements of voltage, current, and temperature for both source and target batteries until the current time-step, each with a length of l. Additionally, y_j^s represents the SOC at the current time-step for sample j. Specifically, each sample comprises previous voltage, current, and temperature measurements from time-step $k - l + 1$ to the current time-step k as input, with the SOC of the current time-step k as the label. This paper aims to establish an SOC estimation model to predict the SOC of the target battery y_j^t utilizing source data X^s and target data X^t, assuming the existence of a distribution discrepancy between the source and target data.

2.2. Deep Neural Network

Our approach leverages a deep neural network architecture to tackle the intricate task of state-of-charge (SOC) estimation. The architecture of our proposed network is depicted in Figure 1, comprising two main modules: a feature extractor and a predictor. The feature extractor plays a pivotal role in capturing the temporal dynamics and patterns inherent in the battery data, facilitating the extraction of informative features crucial for precise SOC estimation. This module may encompass convolutional layers in convolutional neural networks (CNNs), recurrent layers in recurrent neural networks (RNNs), or fully connected layers in feedforward neural networks. Each type of feature extractor possesses distinct strengths and weaknesses, and their performance can vary depending on the specific problem at hand. Subsequently, the extracted features are propagated through a fully-connected layer with a single output node, tasked with mapping these features to the SOC estimation.

2.3. Domain Adaptation

Unsupervised Domain Adaptation (UDA) techniques can be instrumental in mitigating the domain discrepancy between source and target domains. These methods facilitate the alignment of feature distributions across domains, thereby enabling the effective transfer of knowledge from the source domain to enhance state-of-charge (SOC) prediction in the target domain. Metric-based UDA methods aim to alleviate cross-domain distribution discrepancies by applying static criteria. In this study, we leverage a classification-based UDA method, CORAL, and a regression-based UDA method, DARE-GRAM.

Table 1. Results of BiGRU Network for Different Domain Adaptation Methods: Panasonic Battery as Source Domain and LG Battery as Target Domain

Source Temp.	Target Temp.	No TL		CORAL		DARE-GRAM	
		MSE	MAE	MSE	MAE	MSE	MAE
-20°C	-20°C	0.093	0.252	0.103	0.266	**0.031**	**0.143**
	-10°C	0.156	0.342	0.097	0.270	**0.018**	**0.105**
	0°C	0.368	0.522	0.182	0.342	**0.017**	**0.106**
	10°C	0.354	0.513	0.195	0.357	**0.091**	**0.260**
	25°C	0.358	0.516	0.184	0.348	**0.090**	**0.260**
-10°C	-20°C	0.089	0.257	0.025	0.134	**0.018**	**0.110**
	-10°C	0.031	0.148	0.037	0.161	**0.016**	**0.100**
	0°C	0.134	0.305	0.016	0.107	**0.009**	**0.075**
	10°C	0.354	0.513	0.040	0.150	**0.008**	**0.071**
	25°C	0.356	0.514	**0.007**	**0.075**	0.086	0.256
0°C	-20°C	0.049	0.167	0.046	0.162	**0.034**	**0.146**
	-10°C	0.015	0.106	**0.007**	**0.07**	0.008	0.066
	0°C	0.018	0.114	0.013	0.101	**0.005**	**0.057**
	10°C	0.018	0.099	**0.004**	**0.05**	0.011	0.077
	25°C	0.028	0.131	**0.003**	**0.044**	0.02	0.122
10°C	-20°C	0.415	0.56	0.164	0.344	**0.102**	**0.28**
	-10°C	0.376	0.53	0.153	0.324	**0.067**	**0.219**
	0°C	0.366	0.521	0.046	0.192	**0.013**	**0.091**
	10°C	0.03	0.154	0.028	0.151	**0.003**	**0.046**
	25°C	0.009	0.071	**0.005**	**0.065**	0.006	0.068

Correlation Alignment (CORAL) (Sun & Saenko, 2016) stands as a potent domain adaptation technique designed to align the second-order statistics of both the source and target domains. Its primary objective is to diminish the distribution discrepancy between these domains by matching their covariances. This process involves whitening the source and target data to eliminate disparities in variances and subsequently re-coloring the source data to align with the color (covariance) of the target data. By aligning these statistical properties, CORAL effectively enhances the similarity between the source and target distributions, thereby bolstering the transferability of models from the source domain to the target domain. CORAL demonstrates particular efficacy in scenarios where distribution shifts predominantly stem from alterations in data covariances.

DARE-GRAM (Nejjar et al., 2023) harnesses the power of the inverse Gram matrix to align the feature space, taking into consideration the discriminative capability of the final linear layer. This approach prioritizes angle alignment and scale alignment to foster greater compatibility between the source and target domains. The underlying motivation is to identify a feature space conducive to facile learning by a shared linear regressor. Leveraging the ordinary least-squares (OLS) closed-form solution, the method estimates the parameters of the linear layer for regression purposes. By emphasizing the alignment of the angle and scale of the inverse Gram matrix, DARE-GRAM presents a more stable and robust approach compared to direct feature alignment. DARE-GRAM loss function is expressed as follows:

$$L_{DAREGRAM}(F_s, F_t) = \alpha L_{cos}(F_s, F_t) + \gamma L_{scale}(F_s, F_t) \quad (1)$$

where F_s and F_t are extracted features from the source and target domains, respectively. α and γ are hyper-parameters governing the influence of angle and scale alignment, respectively. $L_{cos}(F_s, F_t)$ corresponds to angle alignment, aiming to maximize the cosine similarity between the F_s and F_t. Meanwhile, $L_{scale}(F_s, F_t)$ represents the scaling alignment term, endeavoring to minimize the discrepancy between the k-principal eigenvalues, where k is selected using a specified threshold.

2.4. Training Process

During the training phase, we leverage both source and target data to cultivate domain-invariant representations. The network is guided by two distinct loss functions: the SOC prediction loss, aimed at minimizing the disparity between predicted and actual SOC values in the source domain, and the domain alignment loss, which mandates the resemblance of feature distributions between the source and target domains. The synergy of these loss functions ensures that the network acquires both precise SOC prediction capabilities and domain-invariant features, thereby augmenting SOC estimation accuracy in the target domain. The total loss of the deep network in an end-to-end training scenario is subsequently computed as follows:

$$L_{total} = L_{SOC} + L_{DomainAdaptation} \quad (2)$$

where L_{SOC} denotes the prediction loss, and $L_{DomainAdaptation}$ represents the domain adaptation loss. Since both L_{SOC} and $L_{DomainAdaptation}$ losses are equally critical to the success of the model, we set equal weights to both losses to prevent either loss from dominating. We employ two domain adaptation methods introduced in the previous section to calculate

the domain adaptation loss.

3. Experimental Setup

3.1. Dataset Description

In this study, the efficacy of the proposed method is evaluated using two publicly available LiB datasets: 1) the Panasonic 18650PF dataset (Kollmeyer, 2018) acquired from the University of Wisconsin–Madison, and 2) the LG 18650HG2 dataset (Naguib, Kollmeyer, & Skells, 2020) obtained from McMaster University in Hamilton, Ontario, Canada.

For the Panasonic 18650PF dataset, testing involved brand-new 2.9Ah Panasonic 18650PF cells in an 8 cu.ft. thermal chamber, utilizing a 25 amp, 18 volt Digatron Firing Circuits Universal Battery Tester channel. Similarly, for the LG 18650HG2 dataset, testing was conducted with brand-new 3Ah LG HG2 cells in an 8 cu.ft. thermal chamber, employing a 75 amp, 5 volt Digatron Firing Circuits Universal Battery Tester channel. Both datasets encompassed a series of drive cycles, including US06, HWFET, UDDS, and LA92, performed for each battery. Notably, the battery tests in both datasets were conducted at discrete ambient temperatures ranging from -20°C to 25°C. Figure 2 illustrates the voltage, current, and battery temperature measurements of the two batteries during the LA92 drive cycle.

3.2. Implementation Details

The Panasonic and LG batteries are designated as the "source" and "target" batteries, respectively. Specifically, each experiment involves one Panasonic battery type under a particular ambient temperature serving as the source domain, while LG battery type under a different ambient temperature acts as the target domain. The target data is evenly partitioned into training and testing sets, with the training set utilized for domain adaptation and the testing set employed for performance assessment. As our objective is to assess the efficacy of various unsupervised domain adaptation methods for near-real-time State of Charge (SOC) estimation, we restricted the input sensor data history to the ten most recent observations, ensuring a balanced evaluation across methods without sacrificing generality.

In the deep neural network architecture, we employ Bidirectional Gated Recurrent Unit (BiGRU) modules as feature extractors. GRUs are well-suited for tasks involving sequential information as they efficiently capture temporal dependencies while maintaining a simpler and more streamlined architecture compared to Long Short-Term Memory (LSTM) networks. Moreover, initial experiments conducted as part of our model development phase demonstrated that GRUs outperformed LSTMs in terms of both prediction accuracy and training efficiency. In addition, (Ye & Yu, 2021) demonstrated the efficiency of BiGRU for battery state-of-health

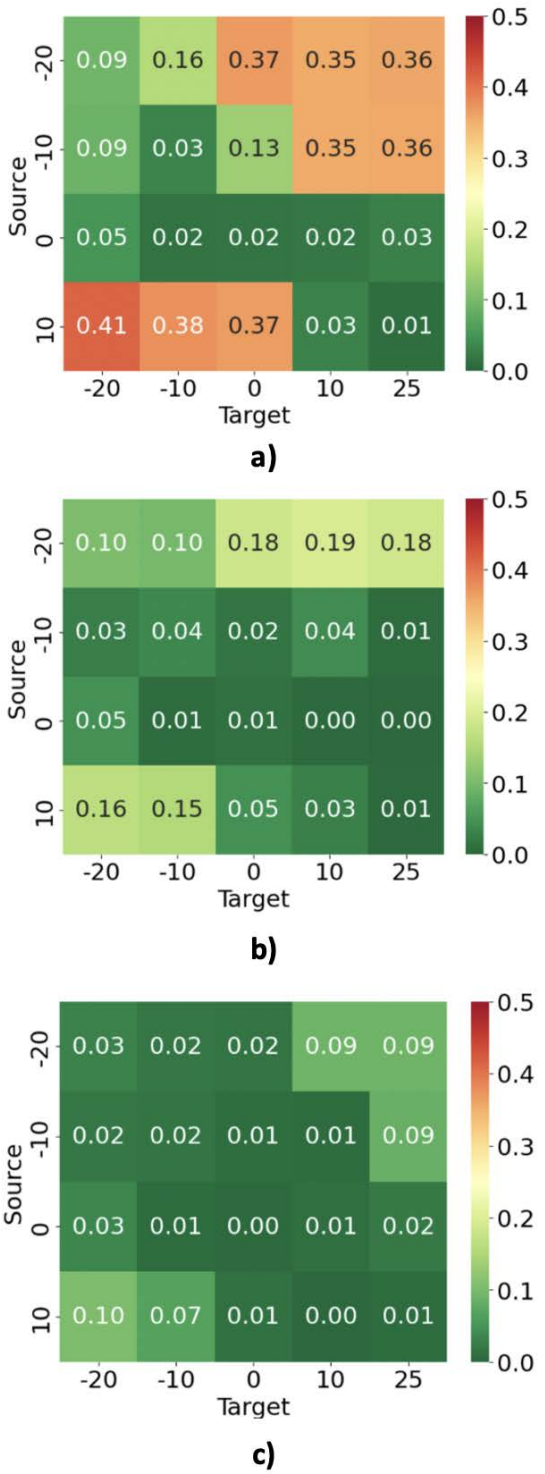


Figure 3. Heatmap of Target Mean Squared Error (MSE) for Different Domain Adaptation Methods under Different Source and Target Temperatures (°C). **a)** No TL, **b)** CORAL, and **c)** DARE-GRAM.

prediction. We run an initial set of experiments to determine the best hyper-parameters to be used in the deep learning

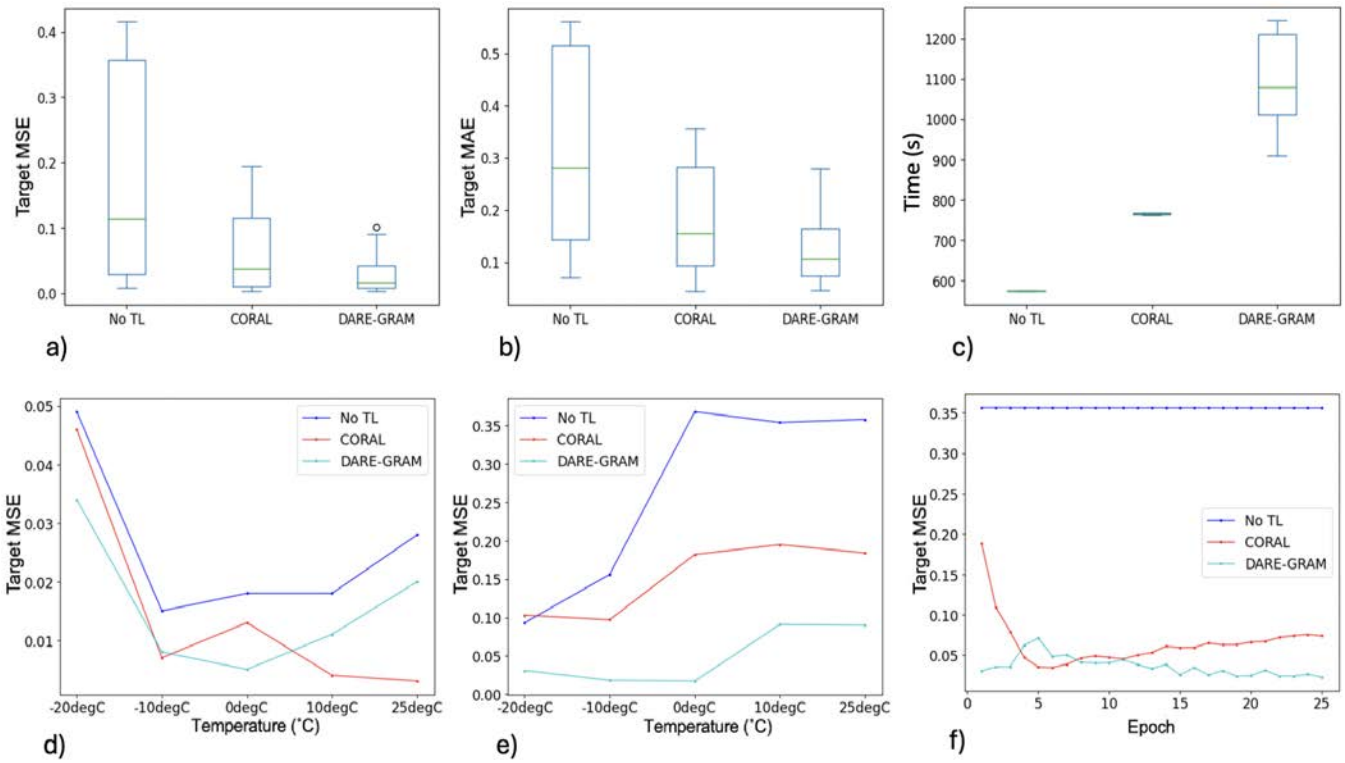


Figure 4. Comparison of Domain Adaptation Methods in State-of-Charge (SOC) Estimation. (a) Target MSE, (b) Target MAE, (c) Training Time (s), (d) Target MSE for Different Temperatures of the Target Battery when the Source Battery is Panasonic in 0°C, (e) Target MSE for Different Temperatures of the Target Battery when the Source Battery is Panasonic in -20°C , and (f) Target MSE Over Training Epochs.

model. The architecture of the feature extractors comprises five layers, each containing 200 hidden units, with a fully-connected (FC) layer consisting of 400 neurons. Additionally, L2 regularization and dropout techniques are applied to enhance the model's generalization and robustness. The number of training epochs is set to 25 for all experiments, with Mean Squared Error (MSE) serving as the loss function for the SOC prediction module.

In addition to utilizing the DARE-GRAM method, we also conduct experiments using the CORAL technique, a well-established classification-based domain adaptation approach, for performance comparison purposes. Furthermore, we perform experiments without any domain adaptation, denoted as the "No TL" model. In "No TL" model experiments, the training process excludes the utilization of target data, with the testing set of target data reserved solely for evaluating the performance of the trained model on the source data. This article utilizes mean-square error (MSE) and mean absolute error (MAE) as the performance evaluation metrics.

4. Results and Discussion

We conduct a thorough analysis of cross-battery state-of-charge (SOC) estimation, comparing the performance of regression-based and classification-based domain adaptation methods. We present the results of our experiments, focusing on the SOC estimation achieved by the BiGRU network with different domain adaptation methods.

Table 1 and Figure 3 summarize the SOC estimation outcomes. In each experiment, the Panasonic battery serves as the source domain, while the LG battery serves as the target domain. While no single domain adaptation method outperforms all others across every experiment, the DARE-GRAM method consistently demonstrates superior performance. Out of the 20 transfer learning tasks conducted, DARE-GRAM outperforms other methods in 15 tasks. Figures 4a,b illustrate box plots for the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values across all tasks for different domain adaptation methods, further highlighting the superiority of the DARE-GRAM approach.

However, it is worth noting that despite its superior performance, the DARE-GRAM method demands more training time. As depicted in Figure 4c, the average training time of the model for DARE-GRAM method is approximately 40% longer compared to other methods. This disparity in computational costs can be attributed to the computation of the domain adaptation loss function. While DARE-GRAM yields

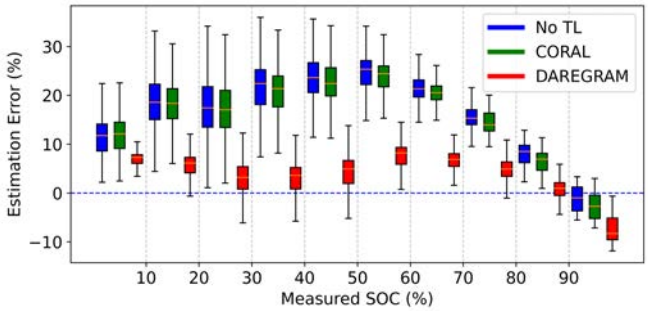


Figure 5. Comparison of Domain Adaptation Methods in SOC Estimation for Panasonic Battery at 10°C to LG Battery at 10°C Task.

impressive results, its computational overhead may pose practical considerations in certain contexts.

While the DARE-GRAM method demands a relatively longer training time over a fixed number of epochs, it exhibits a notably faster convergence, requiring significantly fewer training epochs to reach a stable state. Figure 4f shows the MSE of the target domain for different domain adaptation methods and the "No TL" model over the training epochs for one transfer task (Panasonic -10°C to LG -10°C). This plot shows that with only one epoch, SOC estimations of the target domain for the DARE-GRAM model are significantly more accurate than other methods.

Figure 4d,e shows the results of two different temperatures of the source battery. In each plot, the MSE values of different domain adaptation methods and the "No TL" model are depicted over different temperatures of the target battery. These two plots indicate that cross-battery SOC estimation using the measurements of the source battery under -20°C temperature is significantly more challenging than 0°C. Another important finding is that as the difference between the temperatures of the source and target domains increases, the transfer learning task becomes more rigorous.

Figure 5 illustrates the SOC estimation using different domain adaptation methods for a specific task (Panasonic 10°C to LG 10°C). While the performance of the CORAL method closely resembles that of the "No TL" model, the DARE-GRAM method yields more accurate SOC estimations. DARE-GRAM estimates are somewhat inferior to the other methods when the battery is at full SOC.

The results reveal that for certain transfer tasks, such as Panasonic at 0°C to LG -10°C, even the "No TL" model achieves satisfactory performance. This suggests that at under certain settings, the model trained solely on the source data can effectively estimate the state-of-charge (SOC) for the target data without the utilization of any domain adaptation or transfer learning methods in general.

5. Conclusion

In this work, we introduced a regression-based unsupervised domain adaptation method, DARE-GRAM, for SOC estimation. Through a series of experiments, we assessed the performance and effectiveness of DARE-GRAM in cross-battery SOC estimation, comparing its results with those obtained using the classification-based UDA method, CORAL. Our findings consistently demonstrate the superiority of the DARE-GRAM method in achieving accurate SOC estimation. DARE-GRAM consistently outperformed CORAL, showcasing its robustness and adaptability across various battery domains. Moreover, DARE-GRAM exhibited the ability to prevent negative transfer, ensuring that knowledge transfer did not compromise SOC estimation performance. Furthermore, our results underscored the influence of ambient temperature on model transferability. When the ambient temperatures of both the source and target batteries were similar or closely aligned, the transferability of the model was notably enhanced, leading to improved SOC estimation accuracy. Overall, our study highlights the effectiveness of DARE-GRAM as a powerful tool for enhancing SOC estimation in diverse battery management scenarios, offering valuable insights for future research in the field.

References

Bhattacharjee, A., Verma, A., Mishra, S., & Saha, T. K. (2021). Estimating state of charge for xev batteries using 1d convolutional neural networks and transfer learning. *IEEE Transactions on Vehicular Technology*, *70*(4), 3123–3135.

Bian, C., Yang, S., & Miao, Q. (2020). Cross-domain state-of-charge estimation of li-ion batteries based on deep transfer neural network with multiscale distribution adaptation. *IEEE Transactions on Transportation Electrification*, *7*(3), 1260–1270.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, *22*(14), e49–e57.

Chandra Shekar, A., & Anwar, S. (2019). Real-time state-of-charge estimation via particle swarm optimization on a lithium-ion electrochemical cell model. *Batteries*, *5*(1), 4.

Chen, X., Wang, S., Wang, J., & Long, M. (2021). Representation subspace distance for domain adaptation regression. In *Icml* (pp. 1749–1759).

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, *25*.

How, D. N., Hannan, M., Lipu, M. H., & Ker, P. J. (2019).

State of charge estimation for lithium-ion batteries using model-based and data-driven methods: A review. *Ieee Access*, *7*, 136116–136136.

Hu, J., Hu, J., Lin, H., Li, X., Jiang, C., Qiu, X., & Li, W. (2014). State-of-charge estimation for battery management system using optimized support vector machine for regression. *Journal of Power Sources*, *269*, 682–693.

Khumprom, P., & Yodo, N. (2019). A data-driven predictive prognostic model for lithium-ion batteries based on a deep learning algorithm. *Energies*, *12*(4), 660.

Kollmeyer, P. (2018). *Panasonic 18650pf li-ion battery data.* Mendeley Data, V1. doi: 10.17632/wykht8y7tg.1

Li, Y., Wang, C., & Gong, J. (2016). A combination kalman filter approach for state of charge estimation of lithium-ion battery considering model uncertainty. *Energy*, *109*, 933–946.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning* (pp. 97–105).

Meng, Z., Agyeman, K. A., & Wang, X. (2023). Lithium-ion battery state of charge estimation with adaptability to changing conditions. *IEEE Transactions on Energy Conversion*.

Naguib, M., Kollmeyer, P., & Skells, M. (2020). *Lg 18650hg2 li-ion battery data.* Mendeley Data, V1. doi: 10.17632/b5mj79w5w9.1

Nejjar, I., Wang, Q., & Fink, O. (2023). Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11744–11754).

Ni, Z., Li, B., & Yang, Y. (2023). Deep domain adaptation network for transfer learning of state of charge estimation among batteries. *Journal of Energy Storage*, *61*, 106812.

Oyewole, I., Chehade, A., & Kim, Y. (2022). A controllable deep transfer learning network with multiple domain adaptation for battery state-of-charge estimation. *Applied Energy*, *312*, 118726.

Shen, L., Li, J., Liu, J., Zhu, L., & Shen, H. T. (2022). Temperature adaptive transfer network for cross-domain state-of-charge estimation of li-ion batteries. *IEEE Transactions on Power Electronics*, *38*(3), 3857–3869.

Shen, L., Li, J., Meng, L., Zhu, L., & Shen, H. T. (2023). Transfer learning-based state of charge and state of health estimation for li-ion batteries: A review. *IEEE Transactions on Transportation Electrification*.

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–eccv 2016 workshops: Amsterdam, the netherlands, october 8-10 and 15-16, 2016, proceedings, part iii 14* (pp. 443–450).

Tong, S., Lacap, J. H., & Park, J. W. (2016). Battery state of charge estimation using a load-classifying neural network. *Journal of Energy Storage*, *7*, 236–243.

Wang, Y.-X., Chen, Z., & Zhang, W. (2022). Lithium-ion battery state-of-charge estimation for small target sample sets using the improved gru-based transfer learning. *Energy*, *244*, 123178.

Wang, Z., Feng, G., Zhen, D., Gu, F., & Ball, A. (2021). A review on online state of charge and state of health estimation for lithium-ion batteries in electric vehicles. *Energy Reports*, *7*, 5141–5161.

Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(5), 1–46.

Ye, Z., & Yu, J. (2021). State-of-health estimation for lithium-ion batteries using domain adversarial transfer learning. *IEEE Transactions on Power Electronics*, *37*(3), 3528–3543.

Exploring a Knowledge-Based Approach for Predictive Maintenance of Aircraft Engines: Studying Fault Propagation through Spatial and Topological Component Relationships

Meriem Hafsi [1]

[1] *CESI LINEACT, 15C Av. Albert Einstein, Villeurbanne, 69100, France*
mhafsi@cesi.fr

ABSTRACT

Predictive maintenance has become a highly favored application in Industry 4.0, particularly in complex systems with requirements for reliability, robustness, and performance. Aircraft engines are among these systems, and several studies have been conducted to try to estimate their remaining lifespan. The C-MAPSS dataset provided by NASA has greatly served the scientific community, and several works based on physical models and data-driven approaches have been proposed. However, several limitations related to data quality or data availability of failures persist, and integrating domain knowledge can help address these challenges. In this article, we are currently implementing a new approach based on knowledge coupled with qualitative spatial reasoning to study the propagation of faults within system components until complete shutdown. Region Connection Calculus (RCC8) formal model will be used to describe the component relationships, drawing inspiration from the C-MAPSS dataset.

1. INTRODUCTION

In Industry 4.0, predictive maintenance (PdM) allows for the detection of anomalies and the anticipation of upcoming breakdowns in equipment, machines, or components (Nunes, Santos, & Rocha, 2023). Through the continuous collection of multi-sensor data and system performance analysis, this maintenance strategy relies on machine learning (ML) algorithms capable of building models with the ability to detect early signs of impending failures or malfunctions. Early detection of anomalies allows for prevention, anticipation of corrective actions, and reduced downtime. In this context, PdM solutions rely on estimating the remaining useful life (RUL) before failure (Zio, 2022), which represents the remaining operating time before a component or machine failure. Several approaches are cited in the literature: model-based, data-driven, knowledge-based, or hybrid approaches combining the previous three (Cardoso & Ferreira, 2021). In the aeronotic context, Aircraft engines are among these systems, and several studies have been conducted to try to estimate their remaining lifespan (de Pater, Reijns, & Mitici, 2022). The C-MAPSS dataset provided by NASA [1] has greatly served the scientific community. The solutions proposed in the literature mainly address data-driven approaches (Kumar, 2021; Vollert & Theissler, 2021; Barry, Hafsi, & Mian Qaisar, 2023; Asif et al., 2022), but very few hybrid approaches (Dangut, Jennions, King, & Skaf, 2022) are proposed or tested and no approach attempting to integrate domain knowledge or expert knowledge exists (Barry & Hafsi, 2023; Mayadevi, Martis, Sathyan, & Cohen, 2022).

In this study, we aim to focus on the C-MAPSS dataset referenced in the domain literature and attempt to explore a new approach based on knowledge coupled with qualitative spatial reasoning to study the propagation of faults within system components until complete shutdown. RCC8 rules will be used to describe the component relationships, drawing inspiration from the C-MAPSS dataset (Saxena, Goebel, Simon, & Eklund, 2008), which corresponds to a dataset generated by simulating the operational functioning of aircraft engines, with the aim of evaluating the performance of RUL estimation models. The main objective concerns the modeling of a domain ontology or a semantic graph from the domain knowledge integrated into the C-MAPSS dataset. Spatial and topological representation of system components will be addressed by using RCC8 relations.

2. CONTEXT

Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) developed by NASA, is a simulation tool for a realistic large commercial turbofan engine flights, used for the

[1] data.nasa.gov/dataset/C-MAPSS-Aircraft-Engine-Simulator-Data

Table 1. Overview of C-MAPSS Dataset with segmentation into 4 subsets and description of each subset's characteristics.

	FD001	**FD002**	**FD003**	**FD004**
Train Trajectories	100	260	100	249
Test Trajectories	100	259	100	248
Conditions	1	6	1	6
Failure modes	1	1	2	2

PHM'2008 challenge to generate un large dataset called C-MAPSS dataset (Saxena et al., 2008). It consists of a simulated engine model in the $90,000lb$ thrust class. It also includes an atmospheric model that can simulate operations at various altitudes (sea level to $40,000ft$), Mach numbers (0 to 0.90), and sea-level temperatures (-60 to $103F$). The C-MAPSS dataset is commonly utilized in the field of PdM and engine health prognostic (Vollert & Theissler, 2021). These data are often employed for developing and assessing engine health diagnostic algorithms, failure prediction models, and PdM strategies. The dataset is segmented into different simulation units as demonstrated in Table 1, where each representing an individual engine, with varied failure profiles. Due to its complexity and diversity, C-MAPSS serves as a popular testbed for validating PdM techniques in the aerospace engineering field.

Researchers often leverage this dataset to benchmark their algorithms and methodologies, comparing the performance of different approaches in predicting engine failures and assessing the health status of engines. Moreover, C-MAPSS provides a valuable resource for studying the behavior of engines under various operating conditions and environmental factors (Vollert & Theissler, 2021).

3. Related work

3.1. Predictive Maintenance & PHM Background

In light of the evolving and knowledge-intensive nature of the manufacturing domain, there has been a growing interest in employing semantic technologies (Xia, Zheng, Li, Gao, & Wang, 2022), particularly ontology-based approaches, for PdM. Recent research has introduced various ontologies and rule-based extensions aimed at enhancing knowledge representation and reuse in PdM with several applications in Industry 4.0 (Dalzochio et al., 2020) like in Machinery: (mechanical machines) (Nuñez & Borsato, 2018), (bearings) (Cao, Giustozzi, Zanni-Merk, de Bertrand de Beuvron, & Reich, 2019), elevator running systems (Hou, Qiu, Xue, Wang, & Jiang, 2020), hydraulic systems (Yan et al., 2023), Cyber-Physical Systems (Cao et al., 2022a; Oladapo, Adedeji, Nzenwata, Quoc, & Dada, 2023) and industrial robots (X. Wang, Mingzhou, Liu, Lin, & Xi, 2023). This section provides a review of the most significant research efforts in this area. In

Table 2. Related work applied semantic approaches in the context of Industry 4.0.

Reference	**Application Field**	**Proposition**
(Nuñez & Borsato, 2018)	Mechanical machines	Ontology-based model
(Cao, Giustozzi, et al., 2019)	Bearings / rotating machinery	Ontology-based approach
(Hou et al., 2020)	Elevator running system	Knowledge graph-based approach
(Cao et al., 2022a)	Cyber-Physical Systems	Hybrid approach based on statistical and symbolic AI technologies
(Chhetri, Kurteva, Adigun, & Fensel, 2022)	Hard Drive Failure Prediction	Knowledge Graph Based approach
(Yan et al., 2023)	Hydraulic systems	Knowledge graph-based approach
(X. Wang et al., 2023)	Industrial robots in intelligent manufacturing	PdM method based on data and knowledge
(Oladapo et al., 2023)	Routine maintenance in Industry 4.0	Fuzzified Case-Based Reasoning
(Li, Zhang, Li, Zhou, & Bao, 2023)	steel factory bridge cranes	Knowledge-based approach

(Cao, Samet, Zanni-Merk, De Bertrand de Beuvron, & Reich, 2019), the authors argue that existing PdM approaches have been limited to predicting the timing of machinery failures, while lacking the capability to identify the criticality of the failures. This may lead to inappropriate maintenance plans and strategies. Authors introduce a novel ontology-based approach to facilitate PdM in industry, by combining fuzzy clustering with semantic technologies. Fuzzy clustering techniques are employed to determine the criticality of failures based on historical machine data, while semantic technologies utilize the results of fuzzy clustering to predict the timing and severity of these failures. In (Cao et al., 2022b), the authors address the problem of complexity arising from heterogeneous industrial data, which leads to a semantic gap among manufacturing systems. There is an increasing need for uniform knowledge representation and real-time reasoning in Cyber-Physical Systems (CPS) to automate decision-making processes. In response to this challenge, the authors propose a hybrid approach that combines statistical and symbolic AI. They introduce a system called Knowledge-based System for PdM in Industry 4.0 (KSPMI), which utilizes statistical techniques such as ML and chronicle mining, along with symbolic AI technologies like domain ontologies and logic rules. This hybrid method enables automatic detection of machinery anomalies and prediction of future events. The effectiveness of the approach is demonstrated through evaluation on both real-world and synthetic datasets. In (Chhetri et al., 2022), authors raise the need to improve hard drive failure prediction, given its critical role in computing systems. The

authors point out that existing studies mostly rely on either ML or semantic technology, but each approach has its limitations: ML lacks context-awareness, while semantic technology lacks predictive capabilities. To address these limitations, the authors propose a hybrid approach that combines the strengths of both ML and semantic technology to enhance hard drive failure prediction accuracy. In (Yan et al., 2023), authors are interested in the problems due to the knowledge-intensive and heterogeneous nature of the manufacturing domain, the data and information required for PdM are normally collected from ubiquitous sensing networks. This leads to the gap between massive heterogeneous data/information resources in hydraulic system components and the limited cognitive ability of system users. To address this limitation, the authors propose a virtual knowledge graph-based approach for digitally modeling and intelligently predicting maintenance tasks.

3.2. Knowledge Representation & Spational Reasoning

In the industrial domain, representing knowledge involves organizing and structuring information about processes, systems, and domains. This helps in better understanding and decision-making. With the advancement of technology in Industry 4.0, effective knowledge representation is crucial for optimizing operations and driving innovation. In (Smith et al., 2019), authors highlight the need for a comprehensive ontology to support digital manufacturing, particularly in terms of standardizing terminology across various branches of the advanced manufacturing industries. They propose to develop an upper ontology for the Industrial Ontologies Foundry (IOF), based on the Basic Formal Ontology (BFO), to serve as a foundation for creating a suite of ontologies tailored for digital manufacturing. In (Confalonieri & Guizzardi, 2023) authors discuss the Multiple Roles of Ontologies in Explainable AI. Knowledge-based approaches for RUL estimation have several advantages over other methods (Barry & Hafsi, 2023), including the ability to incorporate domain-specific knowledge and experience into the model, and the ability to handle complex systems where data-driven methods may not be effective. However, they also have limitations, such as being dependent on the availability of expert knowledge and the potential for subjective judgments to influence the model. From an Operations perspective, knowledge-based methods, including fuzzy systems, provide a direct and cost-effective means for RUL estimation by leveraging expert knowledge. These methods prioritize ease of implementation and inter rater reliability. However, their effectiveness is closely tied to the quality of expert input.

Qualitative spatial reasoning, a branch of artificial intelligence, plays a significant role in enhancing decision-making processes within the industrial domain (Fraske, 2022). This approach focuses on analyzing spatial relationships and configurations without precise numerical measurements, allowing for a more intuitive understanding of industrial environments and processes. In the context of Industry 4.0, where smart manufacturing systems heavily rely on interconnected and sensor-rich environments, qualitative spatial reasoning offers valuable insights for optimizing resource allocation, scheduling tasks, and ensuring efficient workflow management (Ladron-de Guevara-Munoz, Alonso-Garcia, de Cozar-Macias, & Blazquez-Parra, 2023).

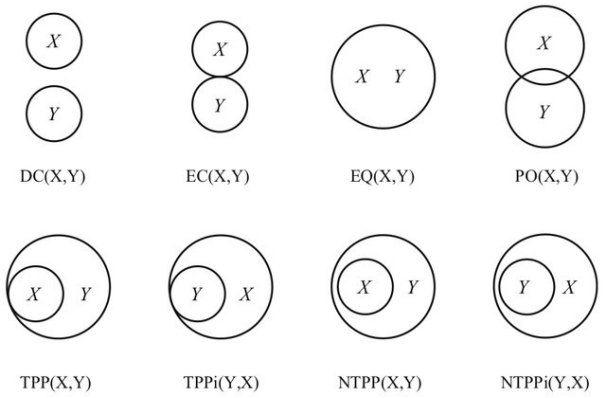


Figure 1. The 8 basic relations of RCC formalism.

RCC (Region Connection Calculus) is a logical formalism used in qualitative geometry intended for representing and reasoning about qualitative spatial relations among regions (Marc-Zwecker, De Bertrand de Beuvron, Zanni-Merk, & Le Ber, 2013). Based on the primitive connection relation $C(x, y)$, where x and y represent spatial regions consisting of a set of points in a plane, delimited by a continuous boundary. The RCC8 formalism defines eight basic relations between regions in space. These relations are exhaustive and mutually disjoint, allowing the definition of any relation between two spatial regions (Y. Wang, Mengling, Liu, & ye, 2018). The eight basic relations are DC (disconnected), EC (externally connected), PO (partially overlapping), EQ (equal), NTPP (non-tangential proper part), TPP (tangential proper part), NTPPi (the inverse of non-tangential proper part), and TPPi (the inverse of tangential proper part) as illustrated in Figure 1 (Lima, Costa, & Moreno, 2019).

4. Proposition

To develop a PdM method based on relationships between industrial components, we propose an approach with application to C-MAPSS aircraft engines as follows: (1) Domain study and advanced data characteristics analysis of aircraft engine components and sensors. (2) Formalization of knowledge in concepts and relationships with a focus on topological relationships between components. (3) Upgrading and describing topological relationships between components based on basic RCC8 relationships. (4) Configuration of a rule-base (based on assumptions) for error propagation across defined

relationships. (5) Determination of alert thresholds for each sensor and component to configure reasoning rules by using data-driven techniques. (6) Test the method and compare it with existing data-driven approaches.

This structured approach describes the steps involved in applying a knowledge-based PdM approach to aircraft engines, integrating domain knowledge with data-driven techniques for effective fault detection and planning main

4.1. Practical Insights into Knowledge Representation for C-MAPSS Scenario

To model and formalize domain knowledge, it is important to understand the functioning of the engine, its components, and the generated data. C-MAPSS consists of four datasets, with each dataset further divided into training and test subsets (Saxena et al., 2008). Each time series originates from a different engine of the same type. Three operational settings, which significantly affect engine performance, are included in the data. Furthermore, the data are contaminated with sensor noise. The engine operates normally at the beginning of each time series but begins to degrade at some point during the series. Multiple aircraft engines undergo varied usage throughout their operational history. A single engine unit may experience different flight conditions from one flight to another. Due to various factors, such as flight duration and environmental conditions, the extent and rate of damage accumulation will vary for each engine. Although the data is simulated, numerous phenomena and challenges have been incorporated to enhance the realism of the dataset. For instance, an initial wear is simulated reflecting typical manufacturing inefficiencies observed in real systems. The initial wear, manifested as minor alterations in pressure, temperature, airflow measurements, etc., is primarily intended to introduce a certain level of manufacturing variability into the data. Indeed, each engine is not identical upon leaving the factory due to manufacturing tolerances and differences in production processes, introducing variability right from the beginning of their use. Additionally, some non-ideal starting conditions or pre-existing degradations are simulated as initial wear due to manufacturing inefficiencies or storage conditions prior to use. Finally, noise is introduced at various stages of the simulation process, ultimately affecting the sensor measurements and mirroring real-world conditions (Saxena et al., 2008).

The engine consists of multiple components, as depicted in Figure 2 (Sánchez-Lasheras, Garcia Nieto, de Cos Juez, Bayón, & González, 2015) :

- **Fan:** The fan component draws in air, providing the initial thrust and airflow into the engine, crucial for combustion.
- **Combustor:** This section mixes fuel with the incoming air and ignites it, generating high-pressure and high-temperature gas for propulsion.

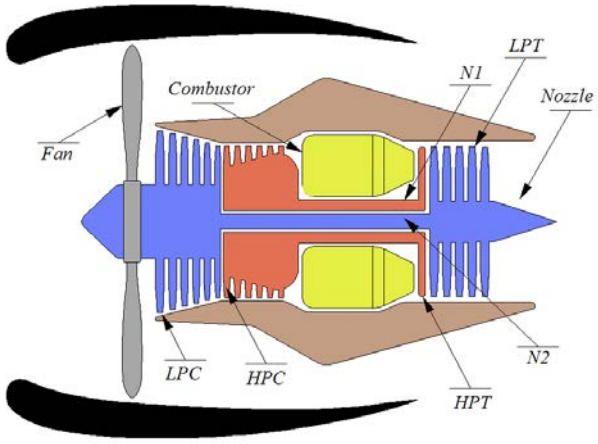


Figure 2. Schematic illustration of an aircraft engine model.

- **LPC (Low-Pressure Compressor):** It further compresses the air before it enters the combustion chamber, enhancing efficiency and power output.
- **HPC (High-Pressure Compressor):** This component significantly raises the pressure of the air, preparing it for combustion and ensuring optimal engine performance.
- **N2:** Represents the low-pressure shaft, connected to the LPC and fan, responsible for driving the fan and low-pressure compressor.
- **HPT (High-Pressure Turbine):** Extracts energy from the high-pressure gas flow to drive the HPC, maintaining compression efficiency.
- **LPT (Low-Pressure Turbine):** Utilizes remaining energy in the gas flow to drive the fan and LPC, contributing to overall engine power generation.
- **Nozzle:** This component accelerates the exhaust gases to produce thrust, directing the flow and converting thermal energy into kinetic energy.

The C-MAPSS dataset simulates engine operation data without providing a detailed description of the sensors utilized. In real-world engines, a diverse array of sensors is commonly employed to monitor various operational and performance parameters. These sensors may encompass:

- **Pressure sensors:** To measure pressure in different parts of the engine, such as combustion chambers, air inlets and outlets, and fuel lines.
- **Temperature sensors:** To monitor temperature in critical areas of the engine, such as combustion chambers, turbines, and exhaust sections.
- **Flow sensors:** To measure the flow rate of fuel, air, or coolant circulating through the engine.
- **Vibration sensors:** To detect abnormal vibrations or signs of imbalance in rotating components of the engine, such as turbine shafts and bearings.

Table 3. Description of the 21 C-MAPSS Sensors.

Sensor ID	Measurement	Unit
T2	Fan inlet temperature	°R
T24	LPC outlet temperature	°R
T30	HPC outlet temperature	°R
T50	LPT outlet temperature	°R
P2	Fan inlet pressure	psia
P15	bypass-duct pressure	psia
P30	HPC outlet pressure	psia
Nf	Physical fan speed	rpm
Nc	Physical core speed	rpm
epr	Engine pressure ratio	-
Ps30	HPC outlet Static pressure	psia
Phi	Ratio of fuel flow to Ps30	pps/psi
NRf	Corrected fan speed	rpm
NRc	Corrected core speed	rpm
BPR	Bypass Ratio	-
htBleed	Bleed Enthalpy	-
Nf_dmd	Demanded fan speed	rpm
PCNfR_dmd	Demanded fan conversion speed	rpm
W31	HPT Coolant air flow	lbm/s
W32	LPT Coolant air flow	lbm/s

- **Speed sensors:** To monitor the rotational speed of engine components, such as turbines and compressors.
- **Position sensors:** To determine the position of valves, flaps, and other moving components of the engine.
- **Exhaust gas sensors:** To analyze exhaust gases and monitor emissions, including gas composition and pollutant levels.

These sensors play a crucial role in collecting engine operation data, which is then used to assess performance, diagnose issues, and predict potential failures as part of PdM and engine health monitoring. Table 3 provides an overview of the sensors included in C-MAPSS.

4.2. Conceptualization and Formalization of Knowledge Domain Ontology

A specialized methodology is used to conceptualize and develop the domain ontology. The Methontology methodology, developed by (Fernández-López, Gomez-Perez, & Juristo, 1997), provides a framework for constructing ontologies at the knowledge level. It includes the identification of the ontology development process and a lifecycle based on evolving prototypes, along with specific techniques for process description as depicted in Figure 3 (Blázquez, Fernández-López, García-Pinar, & Gomez-Perez, 1998).

Our ontology creation follows a systematic approach. The initial step involves the preparation of a formal document meticulously describing the domain to be represented according to the previous section. Subsequently, the conceptualization phase ensues, entailing the definition of concepts, properties, and relationships. For instance, an illustration of the main concepts related by three types of relations in two levels

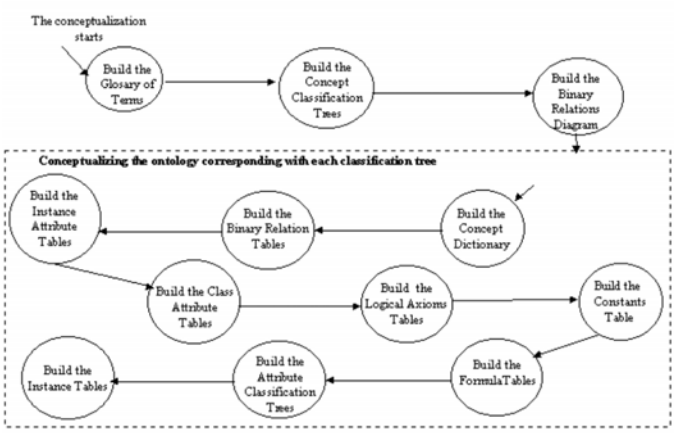


Figure 3. Development process ontology with the Methontology methodology.

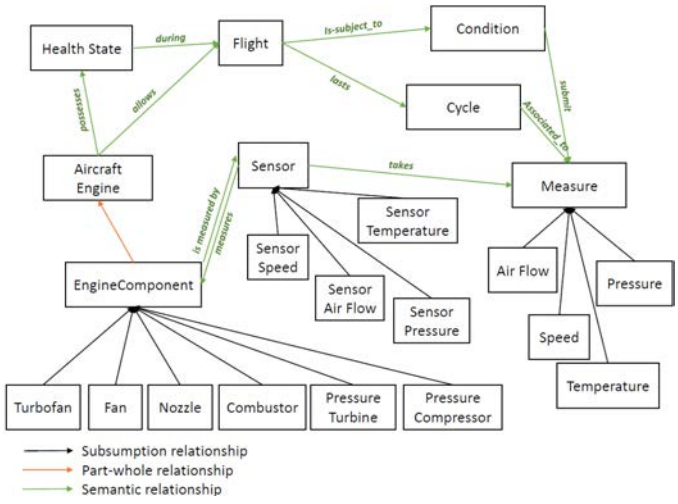


Figure 4. Diagram of ontology classes (concepts) in two main levels.

(subsumption relation/is-a, part-of/part-whole relation, and semantic relations) is provided in Figure 4. Following conceptualization, the third step focuses on formalizing the conceptual knowledge into a language understandable by computers. This modeling can be implemented in an ontology editing tool. In our case, we express the formal ontology in Description Logics (DL) language (Baader, Horrocks, & Sattler, 2005) and implement it using the OWL (Web Ontology Language) format (Taylor, 2009) within the open-source ontology editor Protégé 5.6 [2]. Once the ontology is created, it can be used to annotate and enrich the C-MAPSS dataset with semantic information about the components and their relationships, facilitating advanced analyses and data interoperability. In line with these principles, we establish a conceptual framework to represent pivotal elements and relationships within the C-MAPSS dataset domain. This domain in-

[2]Protégé editor: `https://protege.stanford.edu/`

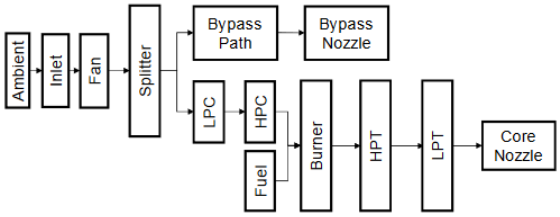


Figure 5. Diagram of modules and their connections in the aircraft engine model.

formation is captured as knowledge within a domain ontology named EngineFailureOntology. Within this ontology, we delineate various concepts, including: Aircraft engine, Engine component, Flight, Condition, Cycle, Measure, Sensor, Heath state, etc. The Definition in DL of Some Design Examples is Provided as Follows:

Engine $\equiv$ ComplexDevice $\wedge$ hasComponent some (Turbine $\sqcap$ Compressor $\sqcap$ Shaft)

Sensor $\equiv$ Device $\wedge$ measures some (Temperature $\sqcap$ Pressure $\sqcap$ Vibration)

TemperatureSensor $\equiv$ Sensor $\wedge$ measures only Temperature

OperationalSetting $\equiv$ Setting $\wedge$ includes some (EngineSpeed $\sqcap$ Load $\sqcap$ AmbientConditions)

4.3. Description of Topological Relationships Between Components

This step involves defining the topological relationships between components using the extension of RCC8 relations and drawing inspiration from diagram in Figure 5, which highlights the interconnections between components.

Some examples to illustrate how RCC8 relations can be used to describe spatial interactions among components of the C-MAPSS engine as follow:

Disjointness DC(Fan,Nozzle) : Fan and Nozzle components are mutually disjoint, as they occupy distinct spatial areas within the engine. This relationship can be expressed by:

$Fan \sqcap Nozzle \equiv \emptyset$

Others disjoitness relationships can be expressed as follows:
$Combustor \sqcap Nozzle \equiv \emptyset$

$Combustor \sqcap Fan \equiv \emptyset$

$LPC \sqcap LPT \equiv \emptyset$

External-Connected EC(HPC, Combustor): The Compressor (HPC) touches the Combustor because the compressed air from the compressor is then directed to the combustor for the combustion process. This relationship can be expressed as follows:

$HPC \sqcap Combustor \neq \emptyset$

Other components are externally connected to each other; these relations can be expressed as follows:

$HPC \sqcap LPC \neq \emptyset$

$LPC \sqcap N2 \neq \emptyset$

$N1 \sqcap Combustor \neq \emptyset$

$HPT \sqcap LPT \neq \emptyset$

$LPT \sqcap N2 \neq \emptyset$

$LPT \sqcap Nozzle \neq \emptyset$

The shaft or rotor (corresponding to the N2 component) is a tangential proper part of the turbine because it is physically attached to the turbine and rotates together with it. Additionally, some of its parts are covered by two other components: N1 and HPT. These relations can be expressed as follows:

$N2 \sqcap HPT \neq \emptyset$

$N2 \sqcap HPC \neq \emptyset$

Partially Overlapping PO(Fan , LPC): The fan overlap some part of the low pressure chamber and it overlap partially, as they share a common space within the engine. This relation can be expressed by:

$Fan \sqcap LPC \neq \emptyset$

The definition of topological relations based on the 2D diagram allows for connecting various components to facilitate the propagation of alerts if a malfunction is observed on a component. This enables the system to identify spatial interactions and dependencies between components, enhancing its capability to detect and propagate alerts effectively throughout the system.

4.4. Reasoning with SWRL rules

Several reasoning rules can be defined in collaboration with domain experts in aeronautics. In this study, we rely on extracting rules from our understanding of the data.

The first rule that can be defined pertains to subjecting a component to significant variations, which may cause fluctuations in sensor values, potentially leading to component fragility and resulting in localized and then generalized malfunction. The risk of impacting neighboring components directly may consequently increase. This rule will be formulated in the form of a SWRL (Semantic Web Rule Language) rule. After defining this rule, the next steps involve loading the time series data from the dataset, initiating the reasoning process, generating a new dataset, and studying the correlation of the new variables obtained through reasoning, in the form of new links or instance values in the knowledge base, with the RUL value. Although it is a logical rule, it is necessary to define

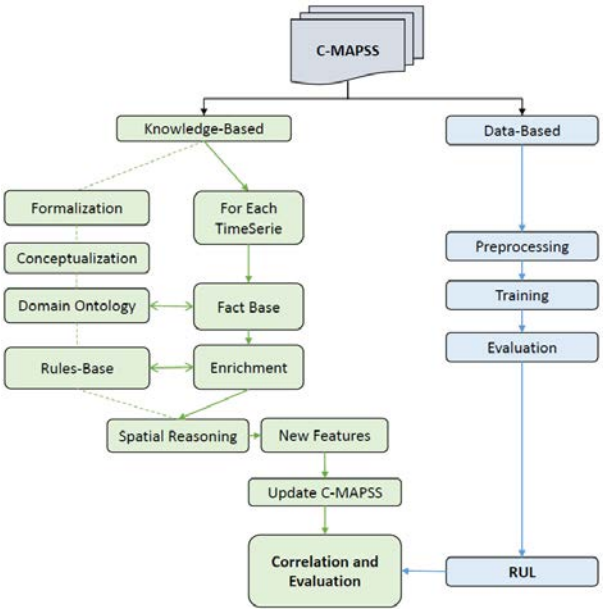


Figure 6. Main steps of the proposed approach involve treating each time series separately and creating features based on the proposed spatial reasoning.

alert thresholds for each sensor to weigh the estimated risk on each component. The definition of these thresholds can be initially done through advanced analysis of the C-MAPSS dataset using ML techniques. Subsequently, collaboration with aeronautical experts can further refine these thresholds.

Finally, we propose this validation technique because the dataset does not provide information on the types of failures and faulty components. Therefore, we will evaluate the validity and performance of our approach by assessing whether the learned knowledge has a positive or negative impact on the estimation of RUL through ML techniques.The process of the proposed approach is illustrated in Figure 6.

5. Discussion

The present study aimed to investigate a novel approach grounded in knowledge representation and spatial reasoning to predict failures by examining fault propagation and its repercussions across components, ultimately impacting the entire system. While similar methodologies have been explored in scientific literature for analogous yet distinct problems, the application of this approach remains novel within the context of our investigation. Despite the inherent complexity associated with its implementation, the potential contribution of this approach towards enhancing the explainability of machine learning (ML) models and elucidating degradation mechanisms holds substantial promise.

5.1. Consensus on the representation of domain and expert knowledge

The conceptualization, formalization, and formulation of rules within this study are predicated upon assumptions crafted within the confines of our research framework. However, it is imperative to acknowledge that such methodologies necessitate close collaboration with domain experts to ascertain the validity and relevance of the defined rules for effective reasoning. To further validate the efficacy of the approach delineated in this article, future endeavors will entail concerted efforts to engage domain experts in refining the formalization of knowledge and iteratively updating the associated reasoning rules. This iterative process of validation and refinement holds the potential to fortify the robustness and applicability of the proposed approach in real-world industrial settings.

5.2. Transition to RCC8 3D Formalism

Furthermore, it is essential to note that the rules of RCC8 pertain to regions in a 2D plane. In this study, we took into account the 2D diagram of components; however, transitioning to 3D objects could offer intriguing avenues for exploration in future research. By extending our analysis to encompass 3D objects, we can potentially enhance the fidelity and accuracy of our predictive models, thereby augmenting the applicability of our approach in diverse industrial scenarios.

5.3. Lack of data on failure types and their origins

Our study is based on the analysis of failure propagation among components, which assumes that a malfunction in one component can be detected or identified. However, the C-MAPSS dataset does not provide the necessary data to obtain this information. Preliminary work is required to estimate the health status of each component and define a threshold indicating failure at its level, as well as to study the propagation to other components. For this purpose, several SWRL reasoning rules can be specified to transition a component to a failure state when its condition is deemed critical. This also involves a detailed analysis of sensor data. For instance, sensors that detect abnormal fluctuations in the data of a component may indicate an impending failure.

6. Conclusion & Future Work

In the context of Industry 4.0 overall, and specifically in the estimation of aircraft engine lifespan, our objective in this article was to investigate the possibility and feasibility of a knowledge-based approach focusing on component degradation as a separate entity before overall system failure, by exploring the potential of qualitative spatial reasoning. The proposed method is currently under implementation, and its results have not yet been evaluated. However, the approach appears to offer tangible benefits, particularly in enhancing our understanding of internal functioning and incident prop-

agation among components. The next steps in this work involve finalizing the proof of concept and obtaining preliminary results. Subsequently, we plan to engage with domain experts to refine the established conceptualization and define reasoning rules that accurately reflect real-world scenarios. Depending on the outcomes, there is potential for applying the method to a cyber-physical system to enhance the explainability of machine learning models in place.

REFERENCES

Asif, O., Haider, S. A., Naqvi, S. R., Zaki, J. F., Kwak, K.-S., & Islam, S. R. (2022). A deep learning model for remaining useful life prediction of aircraft turbofan engine on c-mapss dataset. *IEEE Access*, *10*.

Baader, F., Horrocks, I., & Sattler, U. (2005). Description logics as ontology languages for the semantic web. In D. Hutter & W. Stephan (Eds.), *Mechanizing mathematical reasoning: Essays in honor of jorg h. siekmann on the occasion of his 60th birthday* (pp. 228–248). Berlin, Heidelberg: Springer Berlin Heidelberg.

Barry, I., & Hafsi, M. (2023, December). Towards hybrid predictive maintenance for aircraft engine: Embracing an ontological-data approach. In *20th acs/ieee international conference on computer systems and applications.* Giza, Egypt.

Barry, I., Hafsi, M., & Mian Qaisar, S. (2023). Boosting regression assistive predictive maintenance of the aircraft engine with random-sampling based class balancing. In *13th International Conference on Information Systems and Advanced Technologies.*

Blázquez, M., Fernández-López, M., García-Pinar, J., & Gomez-Perez, A. (1998, 01). Building ontologies at the knowledge level using the ontology design environment.

Cao, Q., Giustozzi, F., Zanni-Merk, C., de Bertrand de Beuvron, F., & Reich, C. (2019). Smart condition monitoring for industry 4.0 manufacturing processes: An ontology-based approach. *Cybernetics and Systems*, *50*(2), 82–96.

Cao, Q., Samet, A., Zanni-Merk, C., De Bertrand de Beuvron, F., & Reich, C. (2019, 01). An ontology-based approach for failure classification in predictive maintenance using fuzzy c-means and swrl rules. *Procedia Computer Science*, *159*, 630-639.

Cao, Q., Zanni-Merk, C., Samet, A., Reich, C., De Bertrand de Beuvron, F., Beckmann, A., & Giannetti, C. (2022a). Kspmi: A knowledge-based system for predictive maintenance in industry 4.0. *Robotics and Computer-Integrated Manufacturing*, *74*, 102281.

Cao, Q., Zanni-Merk, C., Samet, A., Reich, C., De Bertrand de Beuvron, F., Beckmann, A., & Giannetti, C. (2022b, 04). Kspmi: A knowledge-based system for predictive maintenance in industry 4.0. *Robotics and Computer-Integrated Manufacturing*, *74*, 102281.

Cardoso, D., & Ferreira, L. (2021). Application of predictive maintenance concepts using artificial intelligence tools. *Applied Sciences*, *11*(1).

Chhetri, T. R., Kurteva, A., Adigun, J., & Fensel, A. (2022, 01). Knowledge graph based hard drive failure prediction. *Sensors*, *22*, 985.

Confalonieri, R., & Guizzardi, G. (2023). *On the multiple roles of ontologies in explainable ai.*

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, *123*, 103298.

Dangut, M. D., Jennions, I. K., King, S., & Skaf, Z. (2022, mar). A rare failure detection model for aircraft predictive maintenance using a deep hybrid learning approach. *Neural Comput. Appl.*, *35*(4), 2991–3009.

de Pater, I., Reijns, A., & Mitici, M. (2022). Alarm-based predictive maintenance scheduling for aircraft engines with imperfect remaining useful life prognostics. *Reliability Engineering & System Safety*, *221*, 108341.

Fernández-López, M., Gomez-Perez, A., & Juristo, N. (1997, 03). Methontology: from ontological art towards ontological engineering. *Engineering Workshop on Ontological Engineering (AAAI97)*.

Fraske, T. (2022). Industry 4.0 and its geographies: A systematic literature review and the identification of new research avenues. *Digital Geography and Society*, *3*, 100031.

Hou, J., Qiu, R., Xue, J., Wang, C., & Jiang, X.-Q. (2020). Failure prediction of elevator running system based on knowledge graph. In *Proceedings of the 3rd international conference on data science and information technology* (pp. 53–58).

Kumar, K. D. (2021). Remaining useful life prediction of aircraft engines using hybrid model based on artificial intelligence techniques. In *2021 ieee international conference on prognostics and health management (icphm)* (pp. 1–10).

Ladron-de Guevara-Munoz, M. C., Alonso-Garcia, M., de Cozar-Macias, O. D., & Blazquez-Parra, E. B. (2023). The place of descriptive geometry in the face of industry 4.0 challenges. *Symmetry*, *15*(12).

Li, X., Zhang, F., Li, Q., Zhou, B., & Bao, J. (2023). Exploiting a knowledge hypergraph for modeling multinary relations in fault diagnosis reports. *Advanced Engineering Informatics*, *57*, 102084.

Lima, G., Costa, R., & Moreno, M. F. (2019, 11). An introduction to artificial intelligence applied to multimedia. *ArXiv*, *abs/1911.09606*.

Marc-Zwecker, S., De Bertrand de Beuvron, F., Zanni-Merk, C., & Le Ber, F. (2013, 09). Qualitative spatial reason-

ing in rcc8 with owl and swrl..

Mayadevi, B., Martis, D., Sathyan, A., & Cohen, K. (2022, 01). Predictive maintenance of aircraft engines using fuzzy bolt. In (p. 121-128).

Nunes, P., Santos, J., & Rocha, E. (2023). Challenges in predictive maintenance – a review. *CIRP Journal of Manufacturing Science and Technology*, *40*, 53-67.

Nuñez, D. L., & Borsato, M. (2018). Ontoprog: An ontology-based model for implementing prognostics health management in mechanical machines. *Advanced Engineering Informatics*, *38*, 746-759.

Oladapo, K., Adedeji, F., Nzenwata, U., Quoc, B., & Dada, A. (2023, 09). Fuzzified case-based reasoning blockchain framework for predictive maintenance in industry 4.0. In (p. 269-297).

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008, 10). Damage propagation modeling for aircraft engine run-to-failure simulation. *International Conference on Prognostics and Health Management*.

Smith, B., Ameri, F., Cheong, H., Kiritsis, D., Sormaz, D., Will, C., & Otte, N. (2019, 10). A first-order logic formalization of the industrial ontologies foundry signature using basic formal ontology..

Sánchez-Lasheras, F., Garcia Nieto, P. J., de Cos Juez, F., Bayón, R., & González, V. (2015, 03). A hybrid pca-cart-mars-based prognostic approach of the remaining useful life for aircraft engines. *Sensors (Basel, Switzerland)*, *15*, 7062-83.

Taylor, D. (2009, 08). Increasing the expressiveness of owl through procedural attachments. *Focus Symposium Preconference Proceedings*.

Vollert, S., & Theissler, A. (2021, 09). Challenges of machine learning-based rul prognosis: A review on nasa's c-mapss data set..

Wang, X., Mingzhou, L., Liu, C., Lin, L., & Xi, Z. (2023, 08). Data-driven and knowledge-based predictive maintenance method for industrial robots for the production stability of intelligent manufacturing. *Expert Systems with Applications*, *234*, 121136.

Wang, Y., Mengling, Q., Liu, H., & ye, X. (2018, 02). Qualitative spatial reasoning on topological relations by combining the semantic web and constraint satisfaction. *Geo-spatial Information Science*, 1-13.

Xia, L., Zheng, P., Li, X., Gao, R., & Wang, L. (2022). Toward cognitive predictive maintenance: A survey of graph-based approaches. *Journal of Manufacturing Systems*, *64*, 107-120.

Yan, W., Shi, Y., Ji, Z., Sui, Y., Tian, Z., Wang, W., & Cao, Q. (2023). Intelligent predictive maintenance of hydraulic systems based on virtual knowledge graph. *Engineering Applications of Artificial Intelligence*, *126*, 106798.

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*, 108119.

BIOGRAPHIE

Meriem Hafsi was born on 09/10/1990 in Tizi-Ouzou, Algeria. She earned her Master's degree in Project Management in Computer Science from the University of Tizi-Ouzou, followed by a Master's degree in Web Intelligence from the University Jean Monnet of Saint-Étienne. Later, Meriem completed her PhD in Computer Science from Communauté Universités Grenoble-Alpes. Currently, she is a researcher-lecturer at CESI Engineering School and Lineact laboratory in Lyon, where she focuses on research in predictive maintenance in Industry 4.0. Her research interests include utilizing hybrid and innovative approaches to enhance predictive maintenance.

False alarm reduction in railway track quality inspections using machine learning.

Isidro Durazo-Cardenas[1], Bernadin Namoano[2], Andrew Starr[3], Ram Dilip Sala[4], Jichao Lai[5]

[1,2,3,4,5] *Cranfield University, Cranfield, Bedfordshire, MK43 0AL, United Kingdom*

i.s.durazocardenas@cranfield.ac.uk
bernadin.namoano@cranfield.ac.uk
a.starr@cranfield.ac.uk
salaramdilip@gmail.com
damonlai1996@gmail.com

ABSTRACT

Track quality geometry measurements are crucial for the railways' timely maintenance. Regular measurements prevent train delays, passenger discomfort and incidents. However, current fault diagnosis or parameter deviation relies on simple threshold comparison of multiple laser scanners, linear variable differential transformer (LVDT) and camera measurements. Data threshold exceedances enact maintenance actions automatically. However, issues such as measurement error, and sensor failure can result in false positives. Broad localisation resolution prevents trending/ inferencing by comparison with healthy data baseline at the same position over periodic inspections.

False alarms can result in costly ineffective interventions, are hazardous and impact the network availability.

This paper proposes a novel methodology based on convolutional neural network (CNN) technique for detecting and classifying track geometry fault severity automatically. The proposed methodology comprises an automatic flow of data for quality assessment whereby outliers, missing values and misalignment are detected, restored and where appropriate curated. Improved, "clean" datasets were then analysed using a pretrained CNN model. The method was compared with a suite of machine learning algorithms for diagnosis including k-nearest neighbour, support vector machines (SVM), and random forest (RF).

The analysis results of a real track geometry dataset showed that track quality parameters including twist, cant, gauge, and alignment could be effectively diagnosed with an accuracy rate of 97.80% (CNN model). This result represents a remarkable improvement of 38% in comparison with the traditional threshold-based diagnosis. The benefits of this research are not only associated with maintenance intervention cost savings. It also helps prevent unnecessary train speed restrictions arising from misdiagnosis.

1. BACKGROUND

Rail transportation's convenience, punctuality and cost-effectiveness have made it the preferred mode for medium distance travelers and freight (Ghofrani et al., 2018; Wang et al., 2018) Train services as well as the total mileage of track is increasing, which poses a considerable challenge for the effective maintenance of railways infrastructure (Durazo-Cardenas et al., 2018). Degraded rail tracks can cause bumps and swaying when trains pass at high speed and can even cause derailments, putting at risk the safety of passengers. In the event of a failure, delays to the network can also cause significant economic losses (Sasidharan et al., 2020).

Wear and degradation are inevitable, and the railways have implemented safe, tolerance limits for track quality parameters (Railtrack PLC, 1998). Today, tracks are regularly inspected and repaired by dedicated infrastructure maintenance teams. This usually implies a combination of sophisticated track quality inspection trains and on-foot crews that validate and repair the defects flagged by the inspection trains. However, this requires experienced technicians working in hazardous environments, while reducing the availability of the network. Clearly, false alarms raised by the inspection trains contribute to further downtime and costs.

1.1. Measurements and data parameters

The New Measurement Train is an automatic inspection train that is currently the primary method of collecting track geometry data on the British Railways (*New Measurement Train (NMT)*, 2024) . It uses multiple laser scanners, linear variable differential transformer (LVDT), gyroscopes and

accelerometers and cameras Data from thirteen-time domain track parameters are acquired, plus additional imagery systems, accumulating terabytes of data. The NMT can measure track condition at 125mph and cover up to 115,000 miles in a year.

Cross-level, Gauge and Curvature are the essential features of a track alignment, transverse and vertical deviation. Cross-level is defined as the vertical height difference between the tops of the two tracks, while the distance between the two set of rails is known as the Gauge. Curvature is the radius of the arc of the rail, which describes the degree of curvature of the rail. For straight rails, the desired cross-level value approaches zero. while for curved rails, this closely matches the design value.

The *Twist* parameter combines deviations in vertical, and longitudinal dimensions and is typically measured at 3m intervals (Twist3m). The *Cant* parameter describes the difference between the track cross-level and the design cross-level value on curved track.

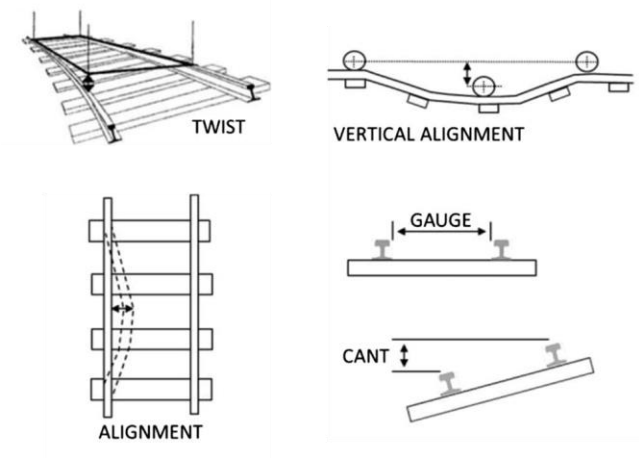


Figure 1 rail track quality parameters. Adapted from D'Angelo et al., (2018).

Top and *Alignment* (AL) account for the deviation between the actual track and the optimal planned path, where top is the vertical distance deviation of the top surface. AL refers to the horizontal distance deviation (Railtrack PLC, 1998). On the other hand, *Dip* is a measure of the depression of the track.

The standard deviation of the measured data parameters is compared to their threshold values for each 1/8 of mile. Exceedances are logged during the train inspection and corrective action notice are issued. Based on the severity of the faults detected and the nominal speed of the line, the health of the track section will be classified as Good, Satisfactory, Poor, Very Poor, and Super Red. Speed restrictions are then issued considering the parameter criticality, with 20 miles per hour being the lowest speed restriction, before track blockage. Network Rail standards(NR/L2/TRK/001/MOD11, 2015 also prescribe the actions to be taken upon threshold exceedances for each parameter, with these ranging from:

1. Block the Line
2. Correct before 36h.
3. Inspect in 72h and correct before 14 to 28 days.
4. Correct before 7 to 14 days.
5. Correct before 14 to 28 days.
6. Add it to the maintenance plan.

1.2. Dataset description

The data used in this this study comprises time series measurement data of the thirteen track parameters described above covering the Southampton-Waterloo line in both directions over a period of one year. This is considered a major line serving many commuter areas including southwestern suburbs of London and the conurbations based around Southampton. Datasets typically comprise CSV acquisition and PDF maintenance team activity logs, and track defect reports. Network Rail reports track quality assessments every 1/8 of mile, with up to 1000 measurements for each parameter acquired. Datasets typically exceed 2 GB.

1.3. Machine learning and related work

Machine learning is often used to analyze large amounts of data and identify connections, offering exceptional potential for anomaly detection analysis (Popov et al., 2022). Recent studies report on machine vision and SVM used to analyze images for track defect detection (Aydin et al., 2021). However, the settings could be significantly costly as it requires high specification tools (cameras, effective fast transmission systems, and efficient storage). Moreover, the large number of images generated bring real time processing challenges hence, affecting on time performance.

Based on time series data, reported accuracy of some traditional machine learning algorithms appears to be relatively low. Considering the disruption, cost and effort involved in railways repairs, higher accuracy is essential. For example, results of SVM algorithm used to detect combined track degradation from car body vibrations reported an accuracy of 80% (Tsunashima, 2019). Lasisi & Attoh-Okine, (2018),used Principal component analysis (PCA) to combine track geometry parameters into a lower-dimensional form and then used SVM, Linear discriminant analysis (LDA), and Random Forest (RF) to detect orbital faults with an accuracy of 92%. However, the true positive rate (precision) is only about 66%, potentially leading to many false alarms.

Several studies have compared the performance of machine learning methods. Sresakoolchai & Kaewunruen, (2019) used a range of supervised and unsupervised machine learning models to analyse track geometry data and sentence faults. The results showed that the non-linear models fitted significantly better, with deep neural network (DNN) having the highest accuracy at 94.3%, followed by convolution neural network (CNN) with 93.8%. The linear models all had

accuracy rates below 50%, with SVM being the poorest at 20%. The results showed that the relationship between features and labels is highly non-linear.

In this context, we propose and effective method to automatically detect and classify track defect using data driven approaches. The next section introduces the methodology.

2. METHODOLOGY

The five-step approach employed in this investigation is illustrated in **Error! Reference source not found.**. The initial data acquisition step is associated with the acquisition of data. Two different types of data were gathered for the analysis. The first datasets comprise of time series signals representing key track quality parameter measurements such as, the location, time, the CANT, the TOP, the gauge, the AL, the Cross-level and the Curvature. The second dataset represents a set of pdf files containing a threshold-based detection report and human based investigation maintenance logs. These files are useful to annotate the track fault observed in the time series data.

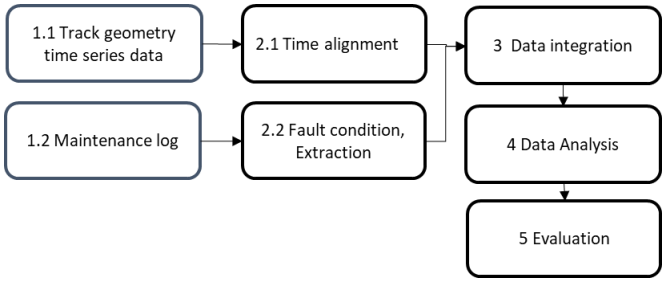


Figure 2 Flow chart of methodology steps.

Step two focuses on the enhancement of individual datasets quality. This was critical step given the issues often encountered with time series data. These datasets typically emanate from a variety of instrumentation sources and are prone to a myriad of consistency issues, including measurement discrepancies caused by instrument malfunctions or user errors. These issues manifest as missing data points, misaligned signals, and a significant presence of noise, each of which can distort the true signal and lead to inaccurate analyses if not properly addressed. To address the absence of data, a localized regression method is implemented. This approach leverages nearby data points to estimate and impute the missing values, assuming that these points observe a similar behavioral pattern. Such assumption is justified given that time series data often exhibits temporal correlation. The efficacy of the imputation process is vital, as it directly affects subsequent analyses. In addition to imputation, the dataset processing phase employs a combination of Dynamic Time Warping (DTW) and Cross Correlation techniques to detect and correct misalignments in the signals. DTW is an algorithm that allows for elastic transformation of time series, enabling the identification of similarities between data sequences that do not align perfectly in time. When used in conjunction with Cross Correlation, it becomes a powerful tool to detect shifts and distortions in the signal, thereby aligning them appropriately for further analysis. Further refining of the dataset includes processing of the maintenance log reports which undergo a procedure to extract and quantify salient features, such as fault type, spatial and temporal coordinates of the occurrences, severity of the detected faults, and the associated maintenance activities required. This information is crucial for understanding the context of the faults and planning preventive measures.

The third stage integrates the outputs of the previous stages, creating a consolidated dataset that is primed for machine learning analytics. This stage is pivotal as it synthesizes the cleaned and aligned time series data with the qualitative information extracted from the maintenance logs, setting the foundation for robust analytical models.

The fourth stage is the heart of the analysis, where two primary categories of machine learning techniques are employed: supervised and unsupervised learning methods. Unsupervised learning techniques, such as the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), are adept at identifying novel fault types that may not have been previously recognized. DBSCAN is privileged for its proficiency in handling noise within the dataset, a common issue in large-scale industrial applications. However, while unsupervised methods excel at detection, they often falter in classification. To counter this, supervised learning methods are applied, harnessing the labelled data produced by unsupervised techniques to train models that can not only detect but also classify fault types. This dual approach ensures that newly occurring faults are not only detected but also categorized correctly. The supervised techniques selected for this stage include robust and widely used algorithms such as Convolutional Neural Networks (CNNs), which are particularly adept at spatial data recognition; k-Nearest Neighbors (kNN), which classifies data based on the proximity to known cases; Random Forests (RF), an ensemble method that improves prediction robustness; and SVMs, which are effective in high-dimensional spaces.

Finally, the last stage focuses on the evaluation of the model's performance, employing three key metrics: precision, recall, and the F1-score. Precision assesses the model's accuracy in predicting fault occurrences, mitigating the risk of false positives. Recall measures the model’s ability to identify all actual fault occurrences, thereby reducing false negatives. The F1-score harmonizes these two metrics, providing a single measure of the model’s accuracy in classification tasks, balancing the trade-off between precision and recall. This comprehensive and iterative process is essential for the identification and classification of faults in complex systems,

such as those encountered in Network Rail's infrastructure. By meticulously processing the data and employing a blend of machine learning techniques, the methodology aims to yield a high-performing model capable of detecting and classifying faults, thus enhancing the maintenance and reliability of the rail network.

3. Results and Discussion

3.1. Dataset analysis

The datasets used in this study come from Network Rail and pertains to track geometry data acquired between 2016 and 2017 during inspections run along the Southampton railway line. These datasets consist of 625 pdfs reports, 300 multivariate time series data which includes track geometry measurements, maintenance team activity logs, and track defect reports. The temporal scope extends over 33 days, with an average monthly coverage of three days. The location parameter, although not a track geometry feature, is a crucial piece of information in the dataset because it provides a baseline for comparing data from different measurement time at the same point. The positioning errors in the recording could reach 100 m. The instrument failure can also cause some positions to be recorded as missing points, resulting in various positions for the same serial number in the data sets. Missing data at a position t is imputed by using and interpolation of different data points observed at a location [t-w] and [t+w], where w (set to 5) helps including of neighbouring data points to enhance the imputation accuracy. For the alignment, the data points must first be aligned so that they are of the same length between data and that data points of the same ordinal number are in the same position. To perform the alignment, first DTW method was used to compute pairwise distance between the measurement, and close distance signal were used to compute a reference signal. Hence maximum value between the cross correlation and the reference signal were used to shift the measurement. **Error! Reference source not found.** and **Error! Reference source not found.** show an example before and after the alignment was performed using the Twist3m measurement.

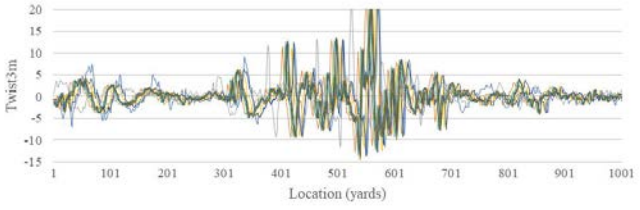


Figure 3 Twist3m measurement before alignment.

After the temporal data alignment, the maintenance log files (shown in **Error! Reference source not found.**) are processed to extract the track identifier (trackid), the mileage, the type of fault (shown in the column “channel”), the peak value observed and the corresponding threshold value. These data are used to locate the fault in the temporal data and hence annotate it. We segmented by file instead of by time series index as the initial experiment on time index split provided imbalanced issues.
The final annotated 300 multivariate time series datasets with about 20000 datapoints each are split into training (60%), validation (20%) and testing (20%).

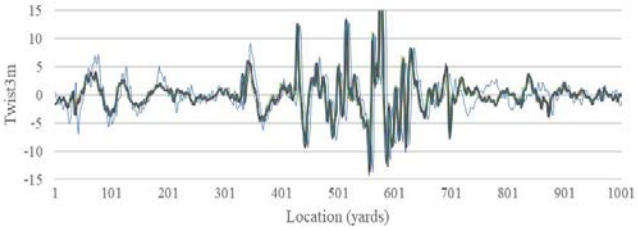


Figure 4 Twist3m after alignment.

Date: 04/07/2016
Train: TRC D
Desc: SOUTHAMPTON<PLATFORM 1>-WATERLOO <FAST LINES>
RST: 0070/2

Route: Wessex
MDU: HW4 Eastleigh MDU
TME: TME Eastleigh
TSM: TSM Eastleigh

ELR	TID	Mileage	GPS	Linespeed	Channel	Peak	Threshold	Action
BML2	3101	078.0926 (42)	5054.4220N:0123.9082W	040	GAUG	16.05	15.00	No Action
BML2	1100	078.0270 (12)	5054.5580N:0123.5278W	085	TW3M	12.25 (1:244)	12.00	No Action
BML2	1100	078.0265 (12)	5054.5590N:0123.5278W	085	GAUG	22.81	15.00	No Action
BML2	1100	078.0230 (10)	5054.5770N:0123.5266W	085	TW3M	13.20 (1:227)	12.00	No Action
BML2	1100	078.0207 (9)	5054.5880N:0123.5254W	085	TL35	-16.92	-16.00	No Action
BML2	1100	078.0173 (8)	5054.6050N:0123.5242W	085	GAUG	-10.76	-5.00	No Action
BML2	1100	078.0149 (7)	5054.6170N:0123.5230W	085	TW3M	14.52 (1:206)	12.00	No Action
BML2	1100	078.0133 (6)	5054.6250N:0123.5218W	085	TW3M	12.58 (1:238)	12.00	No Action
BML2	1100	078.0122 (6)	5054.6300N:0123.5212W	085	GAUG	-5.05	-5.00	No Action
BML1	1100	077.1504 (68)	5054.7020N:0123.5134W	085	TW3M	13.06 (1:229)	12.00	No Action
BML1	1100	077.1498 (68)	5054.7050N:0123.5128W	085	TW3M	-12.64 (1:237)	-12.00	No Action
BML1	1100	077.1494 (68)	5054.7080N:0123.5128W	085	TW3M	12.17 (1:246)	12.00	No Action

Figure 5 Example pdf report.

3.2. Algorithms parameters

The parameters summarised in **Error! Reference source not found.** were empirically tested and configured for this analysis.

For Convolutional Neural Networks (CNN), the setup includes Conv2D, MaxPooling2D, and Dense layers with ReLU and Softmax activations, optimised using adaptive moment estimation (ADAM) with a learning rate of 0.001.

The RF is configured with 100 trees, no maximum depth to allow full growth, a minimum of 2 samples required to split a node and uses the Gini criterion for quality of splits. DBSCAN was set with an epsilon value of 0.5 for maximum neighbourhood distance, and a minimum of 5 samples for core points, while employing Euclidean distance for its metric. DBSCAN being an unsupervised method cannot map automatically with classes, hence we mapped manually the

cluster with a script that computes the cluster centre and the known fault centre.

The k-NN algorithm uses 5 neighbours, uniform weights, automatically selects the algorithm for computing, and utilizes the Minkowski metric. Lastly, the Support Vector Machine (SVM) is configured with an RBF kernel, regularization parameter C set to 1.0, 'scale' for gamma, and a polynomial degree of 3, optimizing for non-linear data separation. Each configuration reflects the algorithm's focus, from spatial clustering and decision forests to similarity-based learning and hyperplane optimization, illustrating the adaptability and specificity required for effective machine learning applications.

Table 1 Methods and parameter settings.

Method	Parameter Configuration Example
CNN	Layers: Conv2D, MaxPooling2D, Dense; Activation: ReLU, Softmax; Optimizer: Adam; Learning Rate: 0.001
RF	Number of Trees: 100; Max Depth: None; Min Samples Split: 2; Criterion: Gini
DBSCAN	Epsilon: 0.5; Min Samples: 5; Metric: Euclidean
kNN	Number of Neighbours: 5; Weights: Uniform; Algorithm: Auto; Metric: Minkowski
SVM	Kernel: RBF; C: 1.0; Gamma: Scale; Degree: 3

3.3. Analysis

The evaluation of the employed method on the testing datasets through precision, recall, and f1-score metrics offers a detailed perspective on their performance in predictive modelling tasks. With a multiclass classification problem, average performance results are computed. As highlighted in **Error! Reference source not found.**, CNN showcased a well-rounded performance with a precision of 97.8%, a recall of 97.69%, and an f1-score of 97.73%, indicating a high degree of accuracy and reliability in identifying relevant instances. RF also demonstrated a strong balance between precision (93.6%) and recall (95.21%), culminating in an f1-score of 94.4%, which underscores its effectiveness in handling various data scenarios. DBSCAN, with a precision of 100%, indicates a perfect identification of relevant instances within its clusters, though its lower recall (88.60%) suggests some relevant instances may not be captured within its clusters, reflected in an f1-score of 93.95%. kNN and SVM both achieved high precision rates (97.6% and 100%, respectively) but with slightly lower recall rates (90.37% and 87.95%, respectively), leading to f1-scores of 93.84% and 93.58%, highlighting their precision in classification but at the expense of some sensitivity.

Table 2 Method performance on the datasets.

Method	Precision **(%)**	**Recall (%)**	**f1-score (%)**
Threshold	59.8	53.80	56.64
CNN	97.8	97.69	97.73
RF	93.6	95.21	94.4
DBSCAN	100	88.60	93.95
kNN	97.6	90.37	93.84
SVM	100	87.95	93.58

Figure 6 CNN confusion matrix.

Focusing on CNN, **Error! Reference source not found.**7 provides a confusion matrix table displaying the performance of a classification model on a testing set, summarizing how well the model distinguishes between five fault classes: Gauge Fault, Twist Fault, AL Fault, Cant, and Top. The matrix shows actual class labels on the vertical axis (TARGET) and predicted labels on the horizontal axis (OUTPUT), with each cell containing the count and percentage of instances. Diagonal cells (in green) represent correctly classified instances, while off-diagonal cells (in red) indicate misclassifications. The overall performance is impressive, with the model correctly classifying 97.80% of the instances (2445 out of 2500) and misclassifying only 2.20% (55 out of 2500). Notably, "AL Fault" has the highest accuracy (99.00% correct), while "Top" has the highest

misclassification rate (4.00%), indicating a specific area for potential improvement. The table also includes sum totals for each class, providing a comprehensive overview of the model's classification capabilities. Figure 8 showing detailed performance tabulation was obtained by computing the relevant performance indicators from Figure 7.

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Gauge Fault	0.98	0.02	0.98	0.02	0.98
Twist Fault	0.98	0.02	0.95	0.05	0.97
AL fault	0.99	0.01	0.98	0.02	0.98
Cant	0.98	0.02	0.99	0.01	0.99
Top	0.96	0.04	0.99	0.01	0.97
Accuracy	0.98				
Misclassification Rate	0.02				
Macro-F1	0.98				
Weighted-F1	0.98				

Figure 7 CNN detailed performance.

Discussion of these results illustrates the inherent trade-offs between precision and recall metrics across different algorithms. CNN and RF, with their balanced precision and recall, are suited for applications where both false positives and false negatives carry significant consequences, providing a robust option for complex classification tasks. The perfect precision of DBSCAN and SVM suggests their utility in scenarios where the cost of false positives is high, making them ideal for applications requiring high confidence in the prediction of positive instances. However, their lower recall rates indicate a potential shortfall in identifying all actual positive instances, which could limit their application in scenarios where missing any positive instance carries a higher risk. kNN, while slightly less precise than SVM or DBSCAN, offers a good compromise between precision and recall, making it a versatile choice for many practical applications. These results underscore the importance of choosing the right algorithm based on the specific requirements and constraints of the task at hand, considering the balance between identifying relevant instances accurately while minimising false identifications. Although these algorithms have various degree of success, they still overperform traditional thresholds technique which precision is 59.8%.

Comparing the machine learning model performance from the literature with the metrics provided reveals a notable advancement in precision, recall, and F1-score. While the literature highlights variances in SVM accuracy from 20% to 80% across applications, this implementation showcases a substantial leap approaching 100% precision for SVM, underscoring a highly effective application or different context. Similarly, The CNN and kNN models not only surpass some of the literature's DNN and CNN benchmarks with CNN achieving a near parity with the highest reported accuracy of 94.3% (Sresakoolchai & Kaewunruen, 2022) but with superior precision and recall. The inclusion of DBSCAN in this analysis, demonstrating a 100% precision, further highlights the potential of selecting and tuning models to suit specific data characteristics and problem contexts. This synthesis underscores the importance of advanced model fine-tuning, the choice of metrics for performance evaluation, and the adaptability of machine learning algorithms to achieve higher efficacy in complex, non-linear problem spaces, especially in critical applications like fault detection in railway systems. In terms of training time, Figure 6 shows the models average estimated time (in blue) and their standard deviation denoted as Std (orange). Random Forest (RF) and Convolutional Neural Network (CNN) methods demonstrate the shortest training times, approximately 1000 and 1200 seconds respectively, with minimal variability. These observations highlight that while SVM is computationally intensive, RF and CNN are more efficient, making them suitable choices when computational resources or time are limited.

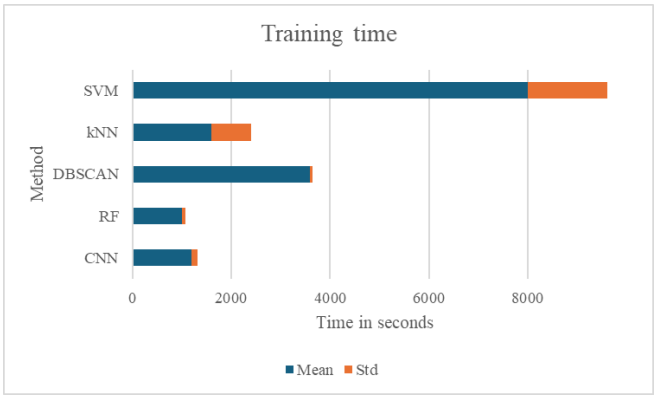


Figure 8 Average Model training time.

4. CONCLUSION

This paper presents a machine learning methodology that successfully improves false alarm rate of railway track quality inspections by 38%. While the datasets examined in this paper only pertain to one specific route, the methods presented here are applicable to all other railway lines across Britain, since the same NMT inspection vehicle is used.

In the railways, repair interventions are costly, typically requiring a manual confirmation of the fault severity, parts,

labour, travel, service disruptions (denial), and penalty fees, as well as being hazardous for the on-foot personnel involved. The ability to correctly diagnose faults also ensures unnecessary speed restrictions are removed, improving journey times and passenger comfort. Evidently, the proposed methodology can potentially have considerable financial impact.

Future work includes an analysis of the potential cost savings achieved using this methodology as well as the integration of context knowledge in the diagnostics.

REFERENCES

Aydin, I., Akin, E., & Karakose, M. (2021). Defect classification based on deep features for railway tracks in sustainable transportation. *Applied Soft Computing*, *111*, 107706. https://doi.org/10.1016/j.asoc.2021.107706

D'Angelo, G., Bressi, S., Giunta, M., Lo Presti, D., & Thom, N. (2018). Novel performance-based technique for predicting maintenance strategy of bitumen stabilised ballast. *Construction and Building Materials*, *161*, 1–8. https://doi.org/10.1016/j.conbuildmat.2017.11.115

Durazo-Cardenas, I., Starr, A., Turner, C. J., Tiwari, A., Kirkwood, L., Bevilacqua, M., Tsourdos, A., Shehab, & Emmanouilidis, C. (2018). An autonomous system for maintenance scheduling data-rich complex infrastructure: Fusing the railways' condition, planning and cost. *Transportation Research Part C: Emerging Technologies*, *89*, 234–253. https://doi.org/10.1016/j.trc.2018.02.010

Ghofrani, F., He, Q., Goverde, R. M. P., & Liu, X. (2018). *Recent applications of big data analytics in railway transportation systems: A survey ☆*. https://doi.org/10.1016/j.trc.2018.03.010

Lasisi, A., & Attoh-Okine, N. (2018). *Principal components analysis and track quality index: A machine learning approach*. https://doi.org/10.1016/j.trc.2018.04.001

New Measurement Train (NMT). (2024). https://www.networkrail.co.uk/running-the-railway/looking-after-the-railway/our-fleet-machines-and-vehicles/new-measurement-train-nmt/

NR/L2/TRK/001/MOD11. (2015). *Inspection & Maintenance of Permanent Way: Track Geometry - Inspections and Minimum Actions*. Network Rail.

Popov, K., De Bold, R., Chai, H. K., Forde, M. C., Ho, C. L., Hyslip, J. P., & Hsu, S. (2022). Big-data driven assessment of railway track and maintenance efficiency using Artificial Neural Networks. *Construction and Building Materials*, *349*, 128786. https://doi.org/10.1016/J.CONBUILDMAT.2022.128786

Railtrack PLC. (1998). Section 8: Track Geometry. In *Track Standards Manual*.

Sasidharan, M., Burrow, M. P. N., & Ghataora, G. S. (2020). *A whole life cycle approach under uncertainty for economically justifiable ballasted railway track maintenance*. https://doi.org/10.1016/j.retrec.2020.100815

Sresakoolchai, J., & Kaewunruen, S. (2022). Railway defect detection based on track geometry using supervised and unsupervised machine learning. *Structural Health Monitoring*, *21*(4), 1757–1767. https://doi.org/10.1177/14759217211044492

Tsunashima, H. (2019). Condition monitoring of railway tracks from car-body vibration using a machine learning technique. *Applied Sciences (Switzerland)*, *9*(13), 2734. https://doi.org/10.3390/APP9132734.

Wang, Y., Correia, G. H. de A., van Arem, B., & Timmermans, H. J. P. (Harry). (2018). Understanding travellers' preferences for different types of trip destination based on mobile internet usage data. *Transportation Research Part C: Emerging Technologies*, *90*, 247–259. https://doi.org/10.1016/J.TRC.2018.03.009

BIOGRAPHIES

Isidro Durazo-Cardenas was born in Hermosillo, Mexico. He is a mechanical engineer with an MSc in advanced automation and design and a PhD in precision engineering from Cranfield University. He is an experienced researcher has led several R&D projects in health monitoring and inspection techniques. Recent research has been on railways autonomy of inspection and repair vehicles. He is a Sr lecturer in Life-cycle engineering at Cranfield University.

Bernadin Namoano is a Lecturer with extensive experience software engineering is currently working on developing novel techniques for asset health management using big time series data. He completed his PhD funded by EPSRC and Unipart-Rail/Instrumentel at Cranfield University in the field of condition monitoring applied to railway assets and was awarded the EPSRC Doctoral Fellowship Prize (2021-2023). His current work involves cognitive digital twins and machine techniques development.

Andrew Starr is chair of maintenance systems at Cranfield University. His work in novel sensing, e-maintenance systems, and decision-making strategies has been recognized with grant support such as H2020 Cleansky 2, iGear (PI), iBearing (PI), EPSRC Platform Grant EP/P019358/1, Through-life performance: From science to instrumentation, and EP/J011630 AUTONOM (PI), AMSCI 31587-233189 CHARIOT (PI), and circa £1.5m in projects for Network Rail in H2020 Shift2Rail.

Ram Dilip Sala received his bachelor's degree in civil engineering from Andhra University, India in 2019. He worked for Amazon until 2022 as Tron associate. He obtained an MSc in Management Information Systems from Cranfield University in 2023.

Jichao Lai is a graduate from Cranfield University (2022) and the Nanjing University of Aeronautics and Astronautics. He has completed an Aerospace Manufacturing MSc. His interests include exoskeleton robotics, iindustrial robot technology and machine learning applications.

Fault Diagnosis of Multiple Components in Complex Mechanical System Using Remote Sensor

Jeongmin Oh[1], Hyunseok Oh[2], Yong Hyun Ryu[3], Kyung-Woo Lee[4], and Dae-Un Sung[5]

[1,2] *School of Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju, 61005, Republic of Korea*
jmoh1010@gm.gist.ac.kr
hsoh@gist.ac.kr

[3,4,5] *Vehicle Performance Degradation Research Laboratory, R&D Center, Hyundai Motor Company, Hwaseong, 18280, Republic of Korea*
skidmarker@hyundai.com
caselee@hyundai.com
dusung@hyundai.com

ABSTRACT

This study proposes an approach to monitor multiple components in complex mechanical systems using a single, externally placed remote sensor. In automobiles and petrochemical plants, where numerous components (e.g., powertrain, bearing, and gear), sensor placement is often compromised by cost and installation environment constraints, resulting in sensing the components far from the regions of interest. To address this challenge, this paper proposes an Operational Transfer Path Analysis (OTPA)-based approach that derives the transfer functions between the vibration excitation source and the measurement point (i.e., receiver). The model for OTPA enables the reverse estimation of the excitation source's signal from the receiver. Subsequently, the estimated (i.e., synthesized) source signal is fed into a diagnostic model to identify system faults. The OTPA and diagnostic models are constructed using neural network architectures, enabling better adaptation to operational conditions and system-induced nonlinearities. The proposed approach is validated from case studies using hydraulic piston pumps in construction vehicles and next-generation electric vehicles.

1. INTRODUCTION

Rotating machine components inevitably produce vibrations during operation, which sensitively reflect the health condition of the rotating machines. Hence, vibration data is predominantly used for fault diagnosis. Such data is often complex and high-dimensional, making effective analysis challenging. Several years ago, deep learning-based approaches were proposed for fault diagnosis in rotating machines. For instance, Zhao, Yan, Chen, Mao, Wang, and Gao (2019) significantly improved the accuracy of motor fault diagnosis using Convolutional Neural Networks (CNN). Their research demonstrates that CNN can successfully classify complex vibration patterns and detect early signs of faults in motors. Additionally, Chen, Zhang, Cao, and Wang (2020) utilized one-dimensional Nonlinear Output Frequency Response Functions and Stacked Denoising Auto-Encoders (SDAE) for diagnosing faults in Permanent Magnet Synchronous Motors (PMSM). The superiority of the proposed method was validated using data from simulations of nonlinear systems such as PMSMs in passenger vehicles.

The existing methods discussed above input data captured from sensors attached to key rotating components such as motors and bearings into deep-learning models for fault diagnosis. Such approaches are effective when sensors can be installed at fault-sensitive points of the rotating machines. In the testbed of the rotating machines, multiple sensors can be attached at various points to diagnose rotating machines effectively using collected vibration data. However, in most operating rotating machines, vibrations arise from numerous rotating parts, and installing sensors at each potential vibration point is often impractical due to cost and environmental constraints. If it is possible to diagnose health conditions using a single sensor installed far from the vibration source, it could significantly reduce costs for data acquisition. Upon further literature review, several innovative studies were found. Choudhary, Mian, Fatima, and Panigrahi (2022) and Yao, Liu, Song, Zhang, and Jiang (2021) creatively proposed diagnosing rotating machines using non-invasive acoustic sensors. Even while acoustic sensors have the amazing benefit of being able to detect signals over long distances, their susceptibility to external

noise can make them inappropriate for use in particular machines such as passenger vehicles.

This paper proposes a novel approach to address two challenges: (1) difficulty in installing vibration sensors on various rotating components within rotating machines, and (2) reduced signal sensitivity and increased interference from external noise when using vibration signals measured at external points. Through Operational Transfer Path Analysis (OTPA), the vibration signal measured at an external point is converted into a vibration signal from the core vibration source, and faults are diagnosed using a denoising deep learning model.

The remainder of this paper is organized as follows: Section 2 describes the OTPA concept, which forms the theoretical background of this paper. Section 3 discusses the proposed approach. Section 4 presents two case studies using a hydraulic pump in a construction vehicle (CV) and a drivetrain of an electric vehicle to verify the proposed approach. Finally, Section 5 summarizes the content of this paper.

2. THEORETICAL BACKGROUND

OTPA is a technique that identifies and quantifies the paths through which vibration and noise are transferred from a source to a receiver (van der Seijs, de Klerk, and Rixen, 2016). Within rotating machines, the input excitation source and the output receiving point can be represented as shown in Figure 1. The relationship between the input excitation source and the output receiving point can be modeled by Frequency Response Function (FRF) as in Eq. (1). through which the response at the receiving point due to an input at the excitation source can be calculated.

$$\mathbf{X}(\omega)\mathbf{H}(\omega) = \mathbf{Y}(\omega) \tag{1}$$

The transfer function is formulated as:

$$\begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_r^{(1)} & \cdots & x_r^{(M)} \end{bmatrix} \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{M1} & \cdots & H_{MN} \end{bmatrix} = \begin{bmatrix} y_1^{(1)} & \cdots & y_1^{(N)} \\ \vdots & \ddots & \vdots \\ y_r^{(1)} & \cdots & y_r^{(N)} \end{bmatrix} \tag{2}$$

where r represents the total count of data sets collected for each point during operation.

In the process of delineating the transfer function matrix, it is imperative to compute the inverse of the matrix X. However, typically, the matrix X is not a square matrix, its inverse matrix cannot be calculated. To overcome this limitation, the Singular Value Decomposition (SVD) technique is employed (Cheng, Zhu, Chen, Song, Zhang, Gao, Liu, Nie, Cao, and Yang, 2022). SVD enables the decomposition of the matrix X as shown in Eq. (3).

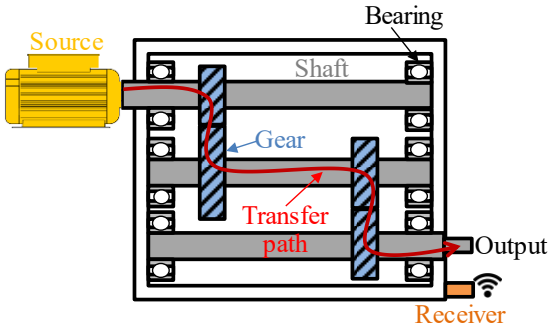
$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{3}$$

where U signifies the unitary matrix; Σ denotes the diagonal matrix constituted by singular values; and $\mathbf{V}^{\mathrm{T}}$ represents the conjugate transpose of the unitary matrix V. Through Eq. (3), it is feasible to obtain the pseudo-inverse of matrix X, a process equivalently captured in Eq. (4).

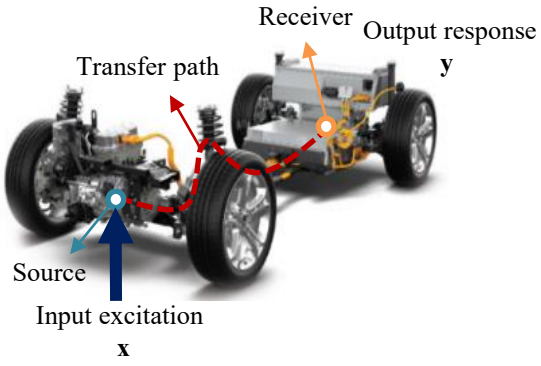
$$\mathbf{X}^{+} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}} \tag{4}$$

Subsequently, the formulation of the transfer function is articulated as:

$$\tilde{\mathbf{H}} = \mathbf{X}^{+}\mathbf{Y} \tag{5}$$



(a)



(b)

Figure 1. Scheme of the transfer path: (a) gearbox, (b) electric vehicle chassis.

3. METHODOLOGY

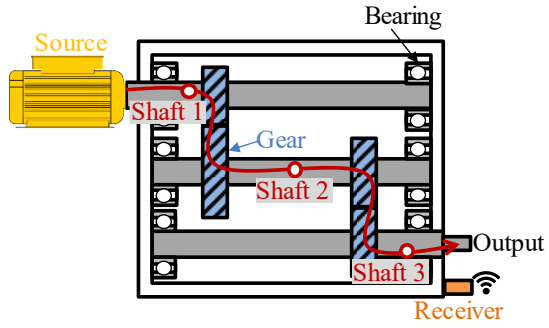
This section presents the proposed approach in this paper. Section 3.1 elaborates on the selection of critical points within rotating machines for vibration signal acquisition and the preprocessing of these signals for the OTPA model. Section 3.2 discusses the architecture of the OTPA model, which transforms measured vibration signals at the receiver into the source vibration signals, and the architecture of the fault diagnosis model that inputs the synthesized vibration

signals. Finally, Section 3.3 describes the training procedure of the models and the metrics employed for performance evaluation.

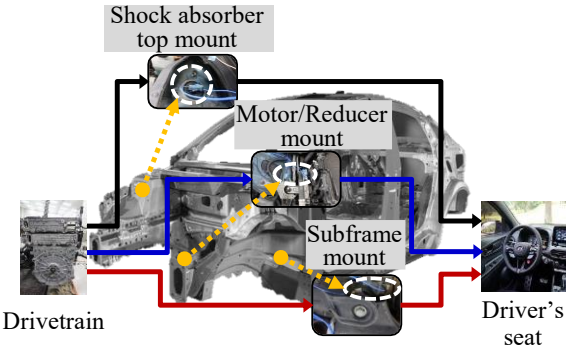
3.1. Data Acquisition and Signal Preprocessing

For the deployment of the OTPA model, initial steps involve the identification of crucial data acquisition points. For instance, in a standard gearbox scenario, sources can be designated as the motor, the receiver as the output case, and transition locations as shafts 1, 2, and 3, as depicted in Figure 2(a). Similarly, for electric vehicles (EVs), the source (excitation point) is determined as the motor, the receiver as the driver's seat, and transition points include the shock absorber mount, subframe mount, and motor/reducer mount, illustrated in Figure 2(b).

Subsequently, acquired vibration signals are converted into frequency spectra using the Fast Fourier Transform (FFT). To mitigate discontinuity issues stemming from finite signal length, a Hanning window is utilized. The overlap rate is 50 %.



(a)



(b)

Figure 2. Source, receiver, and transition locations: (a) CV gearbox, (b) EV chassis.

3.2. Model Building

The proposed model framework, as shown in Figure 3, is comprised of an OTPA deep learning model and a drivetrain fault diagnosis deep learning model. The OTPA deep learning model features two serially connected pairs of a feature extractor and a regressor. The initial OTPA module pair aims to model the transfer function relationship between the receiver signals and those of the transition locations. The subsequent pair focuses on modeling the transfer function relationship between the transition location signals and the excitation source signals.

Feature extractors are composed of 1D convolutional layers, with subsequent regressors mapping the extracted features to vibration signals at transition locations and the drivetrain. Signals measured at the receiver, upon traversing the first feature extractor and regressor combination, are transformed into vibration signals at transition locations; these are then converted into drivetrain vibration signals via the second combination. The final loss function, as represented in Eq. (6), summation of the L_1 losses at transition locations and the drivetrain.

$$L_{total} = L_{trans} + L_{source} \tag{6}$$

where L_{trans} represents the loss at transition locations; and L_{source} denotes the loss at the excitation source. The proposed deep learning model is trained to minimize the combined loss of transition and excitation source locations.

The difference between the proposed deep-learning-based OTPA and conventional OTPA methods lies in the feature extractor. Traditional OTPA methods calculate output signals for specific input signal frequency components using transfer functions, assuming linear independence among measurement directions and locations (de Klerk & Ossipov, 2010). However, such assumptions do not hold when nonlinear interactions within rotating machine components exist. Conversely, the proposed method integrates all frequency components as a singular input to the feature extractor for overall output calculation. Furthermore, the feature extractor leverages activation functions within each layer to learn the nonlinearities between measurement directions and locations. Significantly, 1D convolutional layers facilitate the learning of local band-specific features within the frequency domain, an advantage under variable speed conditions common in rotating machines.

The preprocessing stage involves augmenting the receiver vibration signals with Gaussian noise of varying intensities, creating a dataset with a broad range of signal-to-noise ratios for deep learning model training. While the input data for model training incorporate noise-enhanced signals, the output data utilize the original, unaltered vibration signals. The objective is to train the deep learning model on input data variability, thus ensuring signal transformation accuracy against external noise influences. This approach anticipates effective vibration signal conversion even in harsh operational conditions characterized by significant external noise.

The deep learning model for fault diagnosis employs a conventional Multi-layer Perceptron (MLP) architecture. Input data consist of triaxial (x, y, z) drivetrain frequency domain vibration signals. Parallel-configured MLPs for each channel extract data features, which after layer normalization, are concatenated. The combined features traverse another MLP layer, ultimately outputting indices pertaining to fault types.

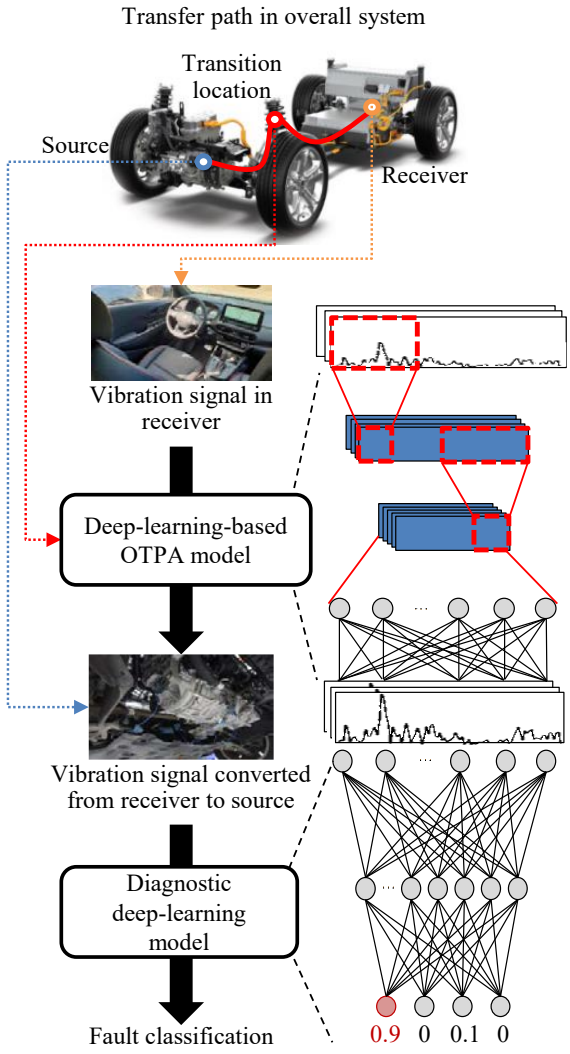


Figure 3. Model construction in the proposed approach.

3.3. Model Training and Performance Evaluation

The synthesized source vibration signals from the deep-learning-based OTPA model are inputted into a trained fault diagnosis deep learning model for assessing the rotating machines' condition. Activation functions across layers in both the OTPA and fault diagnosis models utilize the Rectified Linear Unit (ReLU), with loss functions employing L1 and cross-entropy functions, respectively. The optimizer of choice is Adam.

To objectively evaluate the deep learning models' performance, k-fold cross-validation is conducted. The Mean Absolute Error (MAE) metric assesses the regression fit of the OTPA model. Additionally, accurately realizing local band amplitude fluctuations in the synthesized vibration spectrum, akin to the measured spectrum, is crucial. This aspect can be assessed through variance, thus employing the correlation coefficient as an auxiliary performance metric.

4. EXPERIMENTS

4.1. Case I: Hydraulic Piston Pump

Hydraulic piston pumps are integral in various industries such as construction, maritime, and mining for energy conversion processes. Compared to electric vehicle systems, hydraulic piston pump systems usually have shorter distances between the excitation source and the receiver, resulting in less variance due to external noise. This study investigates the denoising efficacy of the proposed method by artificially introducing external noise to vibration signals obtained at the receiver.

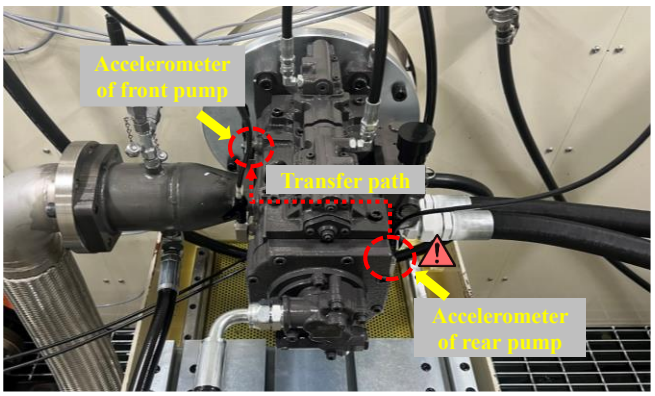


Figure 4. Hydraulic piston pump testbed, measurement points, and transfer path.

4.1.1. Overview

A hydraulic piston pump used in a 22-ton excavator was selected as the subject of our test bed, depicted in **Error! Reference source not found.**. The pump operates at 1800 rpm with a discharge pressure maintained at 350 bar. It is equipped with an axial piston pump at both the front and rear, forming a pair. Artificial defects were introduced into the rear pump piston.

The primary failure modes of axial piston pumps are known to be surface wear between the piston and slipper. Wear occurs due to the chemical degradation of internal lubricants, potentially leading to piston detachment from the slipper, ultimately damaging the piston pump. Based on an

understanding of these failure origins, mechanisms, and modes, piston surface defects were simulated by artificially altering the piston surface tolerances, thereby mimicking increased part tolerances. Tolerances ranging from 0 to 15 μm were classified as normal, 90 to 120 μm as partially-degraded, and 120 to 150 μm as severely-degraded.

Assuming the front pump as the receiver and the rear pump as the excitation source (i.e., the location of fault occurrence), it was hypothesized that utilizing the synthesized vibration signal from the front to the rear pump for diagnosis could achieve higher accuracy. This is because the front pump, physically distant from the actual fault-occurring rear pump, acquires vibration signals that are more susceptible to external noise and contain attenuated signals from the rear pump.

Table 1. Experimental setup for hydraulic piston pump testbed.

Operation condition	1800 rpm
Load condition	350 bar
Oil temperature	50 °C
Number of samples	3 EA
Fault condition (Clearance)	Normal (0~15 μm) Rear partially-degraded (90~120 μm) Rear severely-degraded (120~150 μm)
Sampling rate	10,240 Hz
Measurement points	6 channels (Rear, front x, y, z axis)
Acquisition time	900 sec

Table 2. Train and test dataset configuration for case I.

Training	Original, SNR 20, 10, 0 dB
Test	Original, SNR 20, 15, 10, 5, 0, -5, -10 dB

4.1.2. Data Acquisition and Signal Preprocessing

Vibration data were collected under the conditions detailed in **Error! Reference source not found.** using triaxial accelerometers mounted on the axial piston pump. As described in Section 3.1, the acquired signals were preprocessed and converted into frequency spectra using FFT. A Hanning window of 10,240 data points was used without overlap between consecutive windows. The frequency domain of the transformed vibration signals ranged from 0 Hz to 5,120 Hz. For data set efficiency, frequency domains from 0 Hz to 1,280 Hz were utilized.

To simulate environments with greater distances between vibration acquisition points and more susceptibility to external noise than the hydraulic piston pump samples used in this study, data was augmented by adding Gaussian noise to the assumed receiver signal (front signal), thereby decreasing the SNR as detailed in **Error! Reference source not found.**.

4.1.3. Model Construction and Training

The architecture of the OTPA deep learning model consists of 1D CNN and FC layers, incorporating feature extractors and regressors. Input data comprise frequency spectra of triaxial vibration signals measured at the front pump, while output data consist of the rear pump's triaxial vibration signal spectra. Hyperparameter optimization, conducted empirically, set training epochs at 150, batch size at 64, and learning rate at 0.0001.

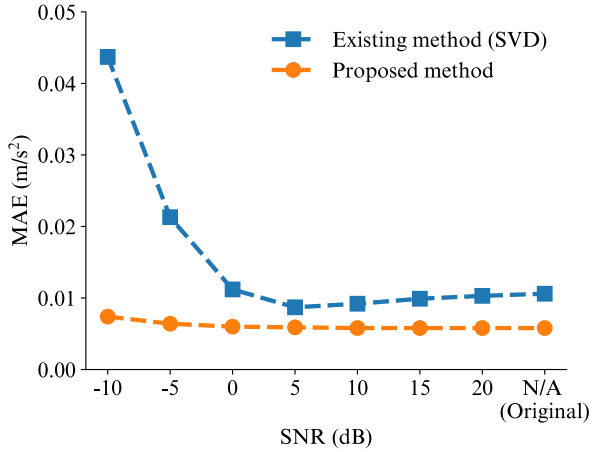
For the fault diagnosis deep learning model, the inputs consist of three-channel excited source spectra with 1,280 data points. Outputs predict the fault vector. Hyperparameter optimization resulted in setting training epochs at 50, batch size at 64, and learning rate at 0.0001. This case study implemented a five-fold cross-validation.

4.1.4. Results

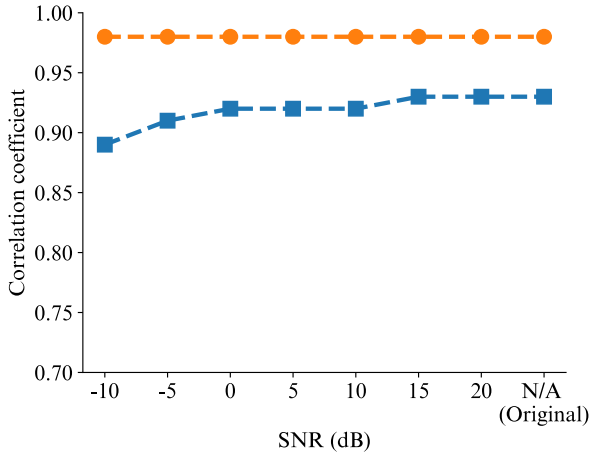
Vibration signal synthesizing results are presented in Figure 5. Using the conventional SVD-based OTPA method, MAE values of 0.0106 m/s² for the original test dataset and 0.0437 m/s² for the dataset with the lowest SNR of -10 dB were observed. In contrast, the proposed method yielded MAE values of 0.0057 m/s² for the original dataset and 0.0070 m/s² for the -10 dB SNR dataset, demonstrating superior performance across all SNR datasets compared to the existing method. Notably, despite all methods showing increasing error trends with higher noise levels, the proposed method exhibited significantly lowered error growth rates, indicating its robustness against noise. Correlation coefficient comparisons also affirmed the superior performance of the proposed method, highlighting the effectiveness of the proposed deep learning-based vibration signal synthesizing method.

Piston fault diagnosis outcomes, comparing front signals to synthesized rear pump signals, are illustrated in Figure 6. Diagnosis performance using front signals was evaluated based on models trained on identical front signals, whereas diagnosis with synthesized rear signals was based on models trained on actual rear signals. Up to SNR 20 dB, both front and synthesized rear signals demonstrated near-perfect diagnostic accuracy. However, as noise levels increased, the diagnostic accuracy based on front signals declined sharply, whereas the accuracy based on synthesized signals remained relatively stable. Notably, for datasets not included in the training process with SNR of -10 dB, front signals showed the diagnostic accuracy of 38.59%, while synthesized signals achieved a significant difference of up to 89.08%. This indicates that, in noisy environments, inputting synthesized

signals from the front to the rear pump enables better differentiation of piston defects, demonstrating the OTPA model's effectiveness in denoising.



(a)



(b)

Figure 5. Performance comparison of the proposed with existing methods for vibration signal conversion: (a) Maximum absolute error and (b) Correlation coefficient.

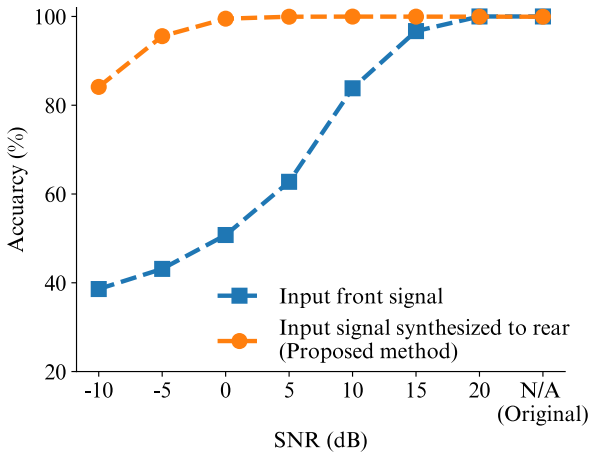


Figure 6. Accuracy comparison of the proposed and existing methods in fault diagnosis.

Table 3. Experimental setup for electric vehicles.

OTPA setup	Sources: electric motor, gearbox Transitions: mounts, subframe, G-bush, knuckle, shock absorber Receiver: driver's seat
Driving conditions	Constant speed of 30, 50, 80, 100 km/h Full acceleration from 0 to 120 km/h with wide open throttle (WOT) 50% acceleration from 0 to 120 km/h with middle tip in (MTI)
Measurements	Sampling rate: 25,600 Hz Data acquisition duration: 25~65 seconds

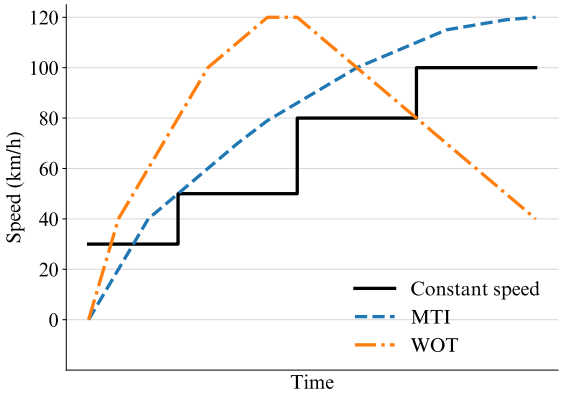


Figure 7. Driving speed profiles.

4.2. Case II: Electric Vehicle Drivetrain

This case study sought to validate the efficacy of the proposed method by implementing it to EVs in the field, synthesizing vibration signals acquired from the driver's seat into drivetrain signals for fault diagnosis. This section consists of subsections including data acquisition and signal

preprocessing, model building, training, and performance evaluation, and results.

4.2.1. Data Acquisition and Signal Preprocessing

Vibration data from an electric vehicle, having completed approximately 300,000 km of operation, were collected. Experimental conditions are detailed in **Error! Reference source not found.**. Using a triaxial accelerometer (DYTRAN 3093M4) with the sampling rate of 25,600 Hz, data were acquired from key components within the EV chassis. The EV drivetrains, comprising motor and gearbox components, were prepared by installing faulty or healthy components to simulate different health condition combinations. Data collection covered various driving conditions, including constant speeds of 30, 50, 80, 100 km/h, and dynamic acceleration-deceleration cycles (Wide Open Throttle; WOT and Middle Tip In; MTI), as illustrated in **Error! Reference source not found.**.

Following the methodology outlined in Section 3.1, collected vibration signals underwent preprocessing before being transformed into frequency spectra using the Hanning window of 25,600 data points and the 50% overlap between consecutive windows. The frequency range of the transformed signals was from 0 to 12,800 Hz. The data from 0 to 4,000 Hz deemed most informative for the dataset.

4.2.2. Model Building, Training and Performance Evaluation

The input to the OTPA deep-learning model consisted of triaxial vibration spectrum data acquired from the driver's seat. These signals were processed through first feature extractor and regressor to convert them into transition locations vibration signals, which were then further converted (or synthesized) into drivetrain vibration signals via second feature extractor and regressor. After optimizing hyperparameters, the training epochs was 150; the batch size was 64; and the learning rate was 0.0001. Performance evaluation of the OTPA model was segmented into constant speeds, acceleration, deceleration, and overall conditions.

The drivetrain fault diagnostic deep-learning model followed the same approach as described in the previous case study. Inputs were the frequency spectra of the converted drivetrain signals, with outputs categorizing into four classes: "normal," "faulty motor," "faulty reducer," and "faulty motor and reducer." After optimizing hyperparameters for the drivetrain fault diagnosis model, the training epochs were 50; the batch size was 64; and the learning rate was 0.0001. The EV drivetrain fault diagnosis model's performance was evaluated under interpolation and extrapolation conditions. In the interpolation condition, models were trained using datasets encompassing all three driving profiles (constant speeds, WOT, MTI), then tested on the same driving conditions. For extrapolation, models were trained exclusively on WOT datasets and tested on constant speeds and MTI datasets not used during training. A three-fold cross-validation was conducted for this case study.

Table 4. Performance evaluation of OTPA models in case of EV.

Transfer path	Driving speed profile	Mean absolute error (dB)		Correlation coefficient	
		Proposed method	SVD	Proposed method	SVD
From driver's seat to transition locations	Constant speed	5.58	6.17	0.77	0.73
	WOT	5.84	6.22	0.67	0.66
	MTI	5.81	6.31	0.70	0.67
	Average (=A)	5.67	6.21	0.74	0.70
From transition locations to drivetrains	Constant speed	5.47	10.25	0.70	0.37
	WOT	6.60	9.16	0.63	0.39
	MTI	6.24	9.89	0.64	0.31
	Average (=B)	5.80	10.02	0.68	0.36
From driver's seat to drivetrains	Total (= A + B)	11.47	16.23	1.42	1.06

4.2.3. Results

The performance of the proposed models was presented in **Error! Reference source not found.**. The proposed model showed the MAE of 11.47 dB and the sum of correlation coefficients of 1.42. The conventional SVD model showed the MAE value of 16.23 dB, significantly higher by 41.4% compared to the proposed model. The sum of correlation coefficients was 25.4% lower at 1.06 than the proposed model. This underscores the superior performance of the proposed model in the operational transfer path analysis.

Subsequent fault diagnosis using the synthesized signals was performed. The performance of the drivetrain fault diagnostic deep-learning model was divided into interpolation and extrapolation conditions for evaluation. The drivetrain fault diagnostic results for each input signal are summarized in Table 5. Interpolation performance showed nearly 100% diagnostic accuracy for both actual driver's seat and

drivetrain signals, and similarly high accuracy for signals synthesized via the proposed methods. However, the conventional SVD approach showed a slight decline to 95.60% accuracy when performing OTPA under varying operational and condition settings (multiple SVDs), and a significant decrease to 34.83% when a single model approach (single SVD) was applied across all conditions.

For extrapolation, the driver's seat signals exhibited a diagnostic accuracy of 90.38%, and drivetrain signals showed 99.32% accuracy. This indicates that the diagnostic model distinguishes drivetrain fault modes more effectively when drivetrain signals are inputted under new driving conditions than when driver's seat signals are used. Additionally, synthesized drivetrain signals, excluding the conventional SVD method, also demonstrated nearly 100% accuracy, similar to actual drivetrain signals. This indicates that the synthesized signals accurately reflect the fault characteristics of actual drivetrain signals, suggesting the effectiveness of the proposed method in accurately diagnosing faults in electric vehicle drivetrains under various operational conditions.

Table 5. Fault diagnostic performance evaluation.

Input	Accuracy (%)	
	Interpolation	Extrapolation
Actual signal measured at driver's seat	99.96	90.38
Actual signal measured at drivetrain	99.98	99.32
Synthesized signal by Single SVD	34.38	33.65
Synthesized signal by multiple SVDs	95.60	93.60
Synthesized signal by proposed method	99.92	99.31

5. CONCLUSION

This paper presented an approach based on Operational Transfer Path Analysis (OTPA) deep-learning modeling for diagnosing faults in rotating machines. The proposed method aimed to address two primary challenges: (1) the absence of vibration sensors at the excited source and transition locations in rotating machines, and (2) the reduction in signal sensitivity due to external noise interference when utilizing vibration signals measured at the receiver. Initially, the transfer paths between the noise generation point (source), transition locations, and the sensor acquisition point (receiver) were defined. Subsequently, a denoising deep-learning model was proposed based on the OTPA concept to capture the nonlinear relationships between the frequency spectra of the excited source, transition locations, and the receiver. To validate the effectiveness of the proposed approach, two case studies were conducted. The case study involving a hydraulic piston pump in construction vehicles demonstrated that the OTPA deep-learning model could convert noise signals into target point vibration signals, significantly reducing noise. At the SNR of -10 dB, the signal conversion accuracy, based on MAE values, indicated that the proposed approach exhibited approximately six times lower error and 50% higher fault diagnosis accuracy than conventional methods. In the case study using an operational electric vehicle, the signal conversion accuracy improved by 59% under various operational conditions compared to the conventional SVD method, and the fault diagnosis accuracy of the proposed approach improved by about 10% under new operational conditions compared to existing seat signal-based diagnostics.

The contributions of this research can be summarized in three key points. First, we proposed an approach that allows for the diagnosis of excited sources and transition locations in rotating machines without installed vibration sensors, using OTPA deep-learning model. Second, the proposed deep-learning model for OTPA demonstrated its capability to address nonlinearities occurring under various operational and fault conditions. Lastly, we introduced a deep learning model that effectively reduces the impact of external noise infiltrating during machine operation and amplifies the excited source's vibration signal.

ACKNOWLEDGMENT

This work was supported by the Hyundai Motor Company and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00144441).

REFERENCES

Chen, L., Zhang, Z., Cao, J., & Wang, X. (2020). A novel method of combining nonlinear frequency spectrum and deep learning for complex system fault diagnosis. *Measurement*, vol. 151, pp. 107190. doi:10.1016/j.measurement.2019.107190

Cheng, W., Zhu, Y., Chen, X., Song, C., Zhang, L., Gao, L., Liu, Y., Nie, Z., Cao, H., & Yang, Y. (2022). AR model-based crosstalk cancellation method for operational transfer path analysis. *Journal of Mechanical Science and Technology*, vol. 36 (3), pp. 1131-1144. doi:10.1007/s12206-022-0206-7

Choudhary, A., Mian, T., Fatima, S., & Panigrahi, B. K. (2022), Deep Transfer Learning Based Fault Diagnosis of Electric Vehicle Motor. *Proceedings of IEEE International Conference on Power Electronics, Drives and Energy Systems*. December 14-17, Jaipur, India. doi:10.1109/PEDES56012.2022.10080274

de Klerk, D., & Ossipov, A. (2010). Operational transfer path analysis: Theory, guidelines and tire noise Application. *Mechanical Systems and Signal*

Processing, vol. 24 (7), pp. 1950-1962. doi:10.1016/j.ymssp.2010.05.009

van der Seijs, M. V., de Klerk, D., & Rixen, D. J. (2016). General framework for transfer path analysis: History, theory and classification of techniques. *Mechanical Systems and Signal Processing*, vol. 68-69, pp. 217-244. doi:10.1016/j.ymssp.2015.08.004

Yao, J., Liu, C., Song, K., Zhang, X., & Jiang, D. (2021). Fault detection of complex planetary gearbox using acoustic signals. *Measurement*, vol. 178, pp. 109428. doi:10.1016/j.measurement.2021.109428

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, vol. 115, pp. 213-237. doi:10.1016/j.ymssp.2018.05.050

BIOGRAPHIES

Jeongmin Oh received the B.S. degree in mechanical design engineering from Pukyong National University, Busan, Republic of Korea, in 2022, the M.S. degree in mechanical engineering from the Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, in 2024, where he is currently pursuing the Ph.D. degree in mechanical engineering. His current research topics include prognostics and health management for rotating machines.

Hyunseok Oh received the B.S. degree in mechanical engineering from Korea University, Seoul, South Korea, in 2004, the M.S. degree in mechanical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2006, and the Ph.D. degree in mechanical engineering from the University of Maryland, College Park, MD, USA, in 2012. He is an Associate Professor with the School of the Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea. His research interests include fault diagnostics, industrial artificial intelligence, and inverse design.

Yong Hyun Ryu received the B.S. degree in mechanical design engineering from Chungang University, Seoul, Republic of Korea, in 2006. His current research topics include the development of harmonic performance for product vehicle and prognostics and health management for electric vehicles in Hyundai Motor Company.

Kyung-Woo Lee received the M.S. degree in 1994, and Ph.D. degree in 2000 in mechanical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea. He has worked for R&D center in Hyundai Motor Company since 2001. His research interests include prognostics and health management for vehicles and future motilities.

Dae-Un Sung received the Ph.D. degree in mechanical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2001. He is a research fellow and the lab. director of Vehicle Performance Degradation Research Lab. In Hyundai Motor Company. His research interests include PHM of power electric system and chassis of mobility vehicles.

Fault Prediction and Estimation of Automotive LiDAR Signals Using Transfer Learning-Based Domain Generalization

Sanghoon Lee[1], Jaewook Lee[2] and Jongsoo Lee[*]

[1]*Reliability and Certification Research Laboratory, Korea Automotive Technology Institute, Chungnam, 31214, Korea*
shlee2@katech.re.kr

[1,2,*]*School of Mechanical Engineering, Yonsei University, Seodaemun-gu, Seoul, 03722, Korea*
shlee2@katech.re.kr
lee1jw@yonsei.ac.kr
jleej@yonsei.ac.kr

ABSTRACT

Autonomous vehicles (AVs) are undergoing level 4 technology development and should have a system that can be operated without driver's intervention, so that it must be possible to diagnose failures and predict life cycle themselves. In this study, we propose a technology to estimate signal changes and sensor faults through transfer learning-based domain generalization (TLDG) using limited actual vehicle test information from LiDAR for autonomous vehicles. Because autonomous vehicles operate in various climate/weather conditions over the world, their mechanical, electrical and electronic components must also have stable performance in all environmental conditions. However, an electronic device, especially laser diode (LD), which is one of core components of LiDAR, shows various degradation aspects depending on environmental conditions. We acquired multivariate LiDAR performance data under various environmental conditions through an actual vehicle test driving of about 2,000 km in summer and winter, and based on this, we create the LiDAR fault diagnosis and performance prediction model generalized to the domain under various environmental conditions. Fault prediction and estimation model created through summer and winter data in the environment domain will also adapt to other environmental conditions such as spring and fall. To develop highly accurate performance estimation models under various environmental conditions based on limited data, it is very important to extract correlations and characteristics between data, including environmental conditions. We employ the data augmentation techniques to solve the problem of lack of training data and apply bi-directional Bayesian transfer learning to generalize data and models under uncertainty. To prove the effectiveness of the present study, the data from actual vehicle tests conducted at different temperatures will be divided into train data and test data, and the validity of the generalized degradation performance estimation model will be statistically validated. The proposed domain generalization method, i.e., TLDG can be utilized to estimate signal changes and sensor faults in LiDAR under unexperienced environmental conditions such as weather changes, and even freezing and hot regions.

1. INTRODUCTION

With the advancement of automobile technology, such as autonomous driving, the need for technology to diagnose and predict automobile failures is emerging. For example, there are AI-based vehicle big data analysis, pre-failure diagnosis and remaining life prediction of parts and systems, and predictive maintenance technology, and these technologies ultimately aim to improve vehicle safety and availability. In level 4 autonomous driving, there is no driver intervention, so the system must independently diagnose and predict failures to ensure safety. Real-time fault diagnosis of autonomous systems is being studied in a variety of ways using data-driven approaches.

AV sensors are composed of composite materials and various electronic elements. As a result, the performance of the sensor may vary depending on the weather environment, and abnormal operation of the sensor may occur under severe conditions (Zhao et al., 2023). Abnormality diagnosis is possible by detecting and scoring abnormal information from changes in sensor performance. Real-time fault diagnosis of autonomous driving systems is being studied in a variety of ways, mainly using edge artificial intelligence (edge AI) and data-based approaches (Gültekin et al., 2022). Research on anomaly detection based on statistical and classification techniques has been active (Ahmed et al., 2016), and recently, research on AI technique-based methods such as adversarial learned denoising shrinkage autoencoder (ALDSAE) has also

been actively conducted in the field of autonomous vehicles (Fang et al., 2023).

Failure prediction and evaluation using transfer learning-based domain generalization presents an innovative approach to solving key problems in the automotive industry and robotics fields (Xu et al., 2019). Domain generalization techniques frequently provide solutions to rotating machinery fault diagnosis problems. The initial approach developed and verified the model under conditions where both training and test data were available. Recently, a method has been proposed to generalize learned knowledge to a new target domain without assuming the availability of test data (Li et al., 2020). Additionally, a new approach called cross-domain augmentation diagnosis, which enables robust defect detection even when the target domain is unknown, is also being studied (Li et al, 2023). However, previous research addresses domain generalization and domain expansion techniques for signals with repetitive operating patterns, such as rotating devices. This is difficult to apply in problems considering multiple stresses because only limited parameters are covered when performing domain generalization. Because automobiles are used all over the world, all systems must operate in a variety of temperature environments. Therefore, when developing safety improvement technologies such as failure prediction, they must be trained or verified considering the distribution of various climate environments. However, due to the considerable time and financial investments required for collecting data through sensor performance measurements across various temperature environments, there are inevitable limitations to the available training and test datasets. Moreover, since sensor components comprise various electronic elements and composite materials, they demonstrate different operating characteristics upon temperature fluctuations. Consequently, there is the potential for poor performance when diagnosing faults based only on limited training data. This problem can be solved through transfer learning-based domain generalization technology, and in this study, we aim to solve the fault detection problem of LiDAR sensors by applying this technology.

In this study, we review the failure modes of LiDAR sensors and select target parameters necessary for failure diagnosis and prediction. The main contributions of the study are as follows:

- Through FMEA analysis of LiDAR defects, we investigate the causal relationship of LiDAR failures that may occur in the vehicle environment, and this allows us to present a practical and versatile model as it deals with hardware level defects by selecting key parameters.
- To solve the problem of accuracy degradation due to outliers encountered in domain generalization problems, we propose an Archimedes spiral-based preprocessing method based on the relationship between input and output data.
- The proposed method provides considerable diversity and flexibility by allowing sensor faults to be predicted with minimal information under diversifying environmental conditions, and the same method can be easily applied for other failure modes.

The rest of the paper is organized as follows. Section 2 gives a failure mode analysis to select key parameters for this study. Section 3 presents preprocessing techniques for outlier data, and Section 4 describes the domain generalization method. The experimentally study is shown in Chapter 5 and finally we conclude in Chapter 6.

2. FAILURE MODE ANALYSIS

LiDAR uses a laser light source to measure distance and recognize the surrounding environment and obstacles. It consists of various components such as laser diodes, thermoelectric elements, signal processing modules, optical lenses, and galvano scanners. In the driving environment, stresses such as heat, vibration, and electrical noise continuously occur, which can causes breakdowns of LiDAR sensors (Chang et al, 2023). The potential failure modes of frequency modulated continuous wave (FMCW) LiDAR are shown in Table 1. Thermal management of FMCW lidar sensors is directly related to sensor performance. When

Table 1. Failure mode analysis of AV LiDAR sensor

Components	Potential failure mode	Potential effects of failure	Potential factors of failure
Laser device	Decrease light intensity	Loss of distance information	High temp/humid., Thermal fatigue
	Fail to detect a returned signal	Increase in false detection	High humid, Vibration
	Fail to keep managed temp.	Non-operation of the sensor	High/Low temperature
Control board	Open circuit	Non-operation of the sensor	Thermal fatigue, Vibration
	Short circuit	Unintended operation	Ingress of dust and moisture
Lens	Fail to focal length	Increase in missed/ false detection	Thermal fatigue, Vibration
	Surface contamination	Increase in missed/false detection	Ingress of dust and moisture
Galvano-meter	Poor responsiveness of actuator	Decrease in sampling rate of scanning	Low temperature, Vibration
	Optical axis misalignment	Increase in missed/false detection	Thermal fatigue, Vibration

analyzing design vulnerabilities through accelerated stress testing, a failure mode in which the laser output of the LiDAR sensor was suddenly cut off under high temperature conditions was identified. This failure mode occurs when the ambient temperature of the laser diode module rises above approximately 75°C. Considering the climate environment of hot weather regions and the sensor self-heating, it can be classified as a failure mode that requires management because it is a condition that can be sufficiently exposed. The laser diode module uses a Peltier-based thermoelectric cooler (TEC) and controls voltage and current to enable the laser diode to maintain a constant temperature. However, when the ambient temperature exceeds a certain range, TEC control ability is lost, and thermal runaway of the laser diode module occurs.

3. OUTLIER DETECTION

In this paper, a study is conducted using the data of the multivariate database acquired through actual vehicle driving test. In the case of actual vehicle driving test data, outliers occur due to uncertainty factors such as road environmental conditions, equipment defects, and frequency errors. To create a robust model, these outliers must be selected and processed in advance and used for training.

As a study on outlier detection, a study has been conducted to propose a fast diagnostic method for internal short circuit (ISC) through local-gravitation outlier detection (Yuan et al, 2023). There is also research on the performance improvement of mechanical failure diagnosis based on audio signal analysis (MFDA) based on outlier detection (Tang et al, 2022).

We would like to propose a signal processing method through Archimedes spiral as a new outlier detection method. Using the following method, Archimedes spiral can be used to create a deep learning model that is generalized to the domain without overfitting. When Archimedes Spiral is expressed using a polar coordinate system, it can be expressed as Equations (1) using the real constant a, b and the angle θ. Changing the parameter b controls the distance between loops. Since we want to check that the output value changes for the input variable, we just need to check how the distance of the spiral changes at a specific angle set as the input variable. Here, when input data is x_{input} and output data is y_{output}, it can be expressed as Equation (2). Next, to solve the problem that is not visually clear by setting the starting point as the origin when drawing the spiral, the data was normalized between 2π and 4π at the start of the second spiral, as shown in Equation (3) and (4). When the data is sorted through this, data including the uncertainty factor appears as an area, and data processing based on the confidence interval is possible according to the data area.

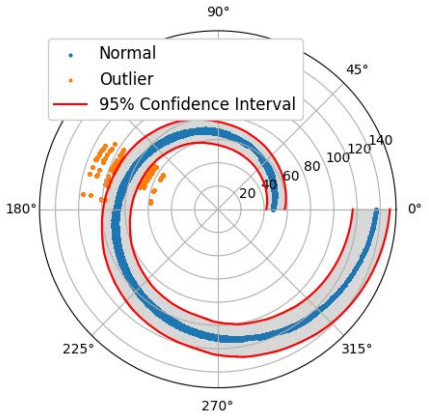
$$r = a + b \cdot \theta \tag{1}$$

$$(r, \theta) = (x_{input} \cdot y_{output}, x_{input}) \tag{2}$$

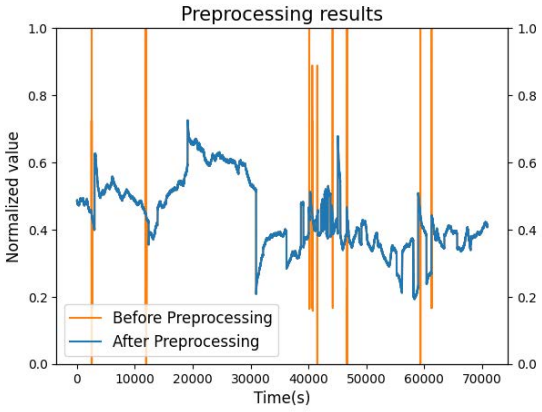
$$x_{normalized} = \frac{x_{raw} - x_{min}}{x_{max} - x_{min}} \tag{3}$$

$$\begin{aligned} x_{input} &= 2\pi \cdot x_{normalized} + 2\pi \\ y_{output} &= 2\pi \cdot y_{normalized} + 2\pi \end{aligned} \tag{4}$$

As illustrated in Figure 1, if you enter the temperature data of LiDAR in x and the current TEC current to be estimated in y, normal data and outlier data are distinguished. In addition, it is possible to statistically process the outlier of the data through the confidence interval. Ultimately, we want to create a regression deep learning model through the data classified as normal and generalize it to the domain.



(a)



(b)

Figure 1. Archimedes spiral of LiDAR temperature to TEC current of laser diode, (a) Archimedes spiral, (b) Comparison of preprocessing results

4. Transfer Learning-Based Domain Generalization

In this paper, we propose a transfer learning-based domain generalization method to overcome the limitations of data acquisition under various environmental conditions, including extreme conditions. It is known that transfer learning can improve predictive performance in terms of interpolation or extrapolation of the model by utilizing only a small amount of data from the target domain based on the model generated using the source domain (Weiss et al, 2016). However, domain generalization differs in predicting physical phenomena in the unseen domain region using improved models. We intend to gradually transfer a small amount of data to the target domain, use it to predict data in the new unseen area, and use it again as target data for transfer learning to generalize the domain. As illustrated in Figure 2, First, we created a regression model based on deep natural network (DNN) that predicts TEC current using temperature, humidity, current, and voltage data for the underlying source domain. Next, after importing the feature extraction area of the underlying source domain model to be untrainable, transfer learning was performed by adding new layers for transfer learning.

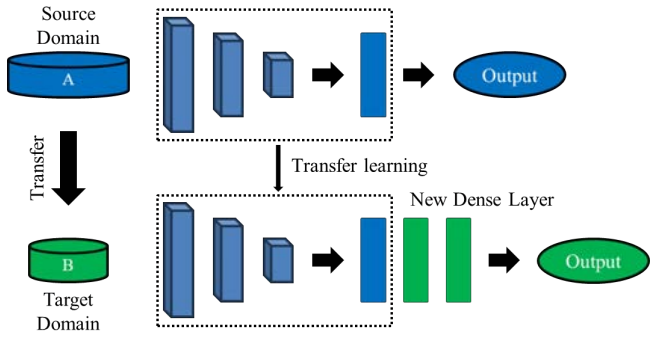


Figure 2. Architecture of transfer learning

5. Experimental Study

5.1. Datasets

In this paper, LiDAR data acquired through actual vehicle driving data in summer were used. Temperature, humidity, current, and voltage data were measured including TEC current of the Laser diode, and actual vehicle driving test data of more than 1000 km including city, country, and highway were acquired. The sensor used was the FMCW 4D LiDAR G-Series from Infoworks, and the internal and external temperatures of the sensor were measured using the SHT45-AD1B temperature sensor from Sensirion. Actual vehicle test data was obtained by installing the LiDAR on a Hyundai Azera and collecting TEC current and temperature data under actual driving conditions. Through this, 283,706 data points were acquired every 0.25 seconds. As illustrated in Figure 3, The left axis represents temperature, and the right axis represents TEC current. The environmental temperature is 28.96°C to 45.65°C, and in the case of TEC current, it may be confirmed that outlier exists. Among them, 268,402 data in which outliers were removed were selected through outlier detection based on Archimedes spiral. In addition, the ratio of training, validation, and test data was divided into 0.6, 0.2 and 0.2, and normalization was performed and used for training and model evaluation. For training, the temperature range of the training data and the temperature range of the test data were set by setting the scenarios of interpolation and extrapolation, respectively. And for transfer learning, 0.5% of the test data was arbitrarily extracted and set as data from the target domain.

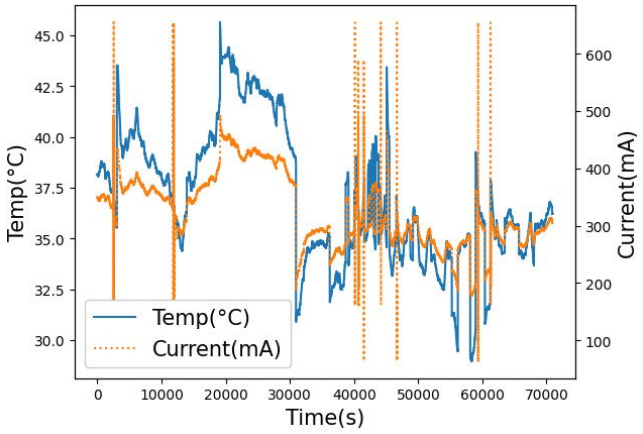


Figure 3. Actual driving test data for temperature (°C) and TEC current (mA)

5.2. Results

This paper proposes a method for predicting data based on outlier detection and transfer learning using actual driving data. All data were utilized in a state where outlier detection was conducted by Archimedes spiral. This method shows superior performance compared to other preprocessing techniques. Specifically, when comparing accuracy using methods interquartile range (IQR) and Hampel filter, the outlier detection performance was 3.23% for method IQR and 83.26% for method Hampel filter, while our proposed method demonstrated a performance of 100%. Before represents the result before performing transfer learning, and after represents the performance after transfer learning. First, looking at interpolation case 1, data from 35°C to 40°C were used as test data, and other data were used as training data. Although the error improved from 0.01 to 0.0009 based on mean absolute error (MAE), the r-squared of the DNN model was so good that the interpolation problem did not require transfer learning. This was also shown in the case of interpolation case 2. However, in the case of extrapolation, the performance error of the model before transfer learning is relatively large. However, if improvement is made through transfer learning, in case 3, it improved from 0.81 to 0.96 based on r-squared, and MAE also improved from 0.027 to

Table 2. Comparison table of transfer learning results

Scenarios	Case	Train data (°C)	Test data (°C)		R-squared	MAE
Interpolation	Case 1	28.96 ~ 35.00, 40.00 ~ 45.65	35.00 ~ 40.00	Before	0.98	0.010
				After	0.98	0.009
	Case 2	28.96 ~ 33.00, 43.00 ~ 45.65	33.00 ~ 43.00	Before	0.97	0.021
				After	0.99	0.011
Extrapolation	Case 3	28.96 ~ 40.00	40.00 ~ 45.65	Before	0.81	0.027
				After	0.96	0.010
	Case 4	28.96 ~ 35.00	35.00 ~ 45.65	Before	0.77	0.066
				After	0.99	0.014

0.010. Finally, in the case of case 4, which used only data up to 35°C as training data, it improved from 0.77 to 0.99 based on r-squared, and MAE also improved from 0.066 to 0.014. The results of case 4 are expressed visually through Figure 4. The prediction accuracy gradually decreases in the case of test data far from the area of the train data. However, after transfer learning, prediction accuracy has improved even in areas away from training data.

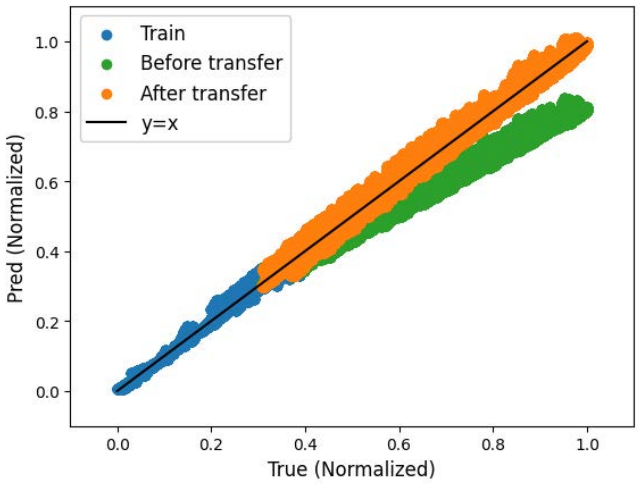


Figure 4. Extrapolation result for Case 4 in Table 2

6. Conclusion

In this paper, we propose a transfer learning-based domain generalization model for FMCW LiDAR signals that change with external temperature changes. This introduces a new approach to predicting LiDAR sensor errors by allowing sensor behavior to be predicted in unseen regions. LiDAR failure mode analysis justifies the selection of TEC current as a predictor and experimentally demonstrates the nature and validity of this signal. Real-world driving data often contains outliers due to various errors, and using Archimedes Spiral-based data preprocessing improves the prediction accuracy of the model. In the generalization task, temperature, humidity, current, and voltage data from the source domain were used, and transfer learning was performed using a DNN-based regression model and a new Dense Layer. The generalized model showed high accuracy and proved to be effective for extrapolation. Extensive training data covering a variety of climate conditions can further improve the accuracy of this model. The existing model was developed using only summer data, but future iterations will incorporate winter data to develop a domain generalized model that takes low-temperature environments into account. Through interpolation methods, it may be possible to predict sensor failure under all climatic conditions in Korea. Our goal is failure prediction under severe weather conditions. This is an extrapolation technique, and we plan to develop a domain-generalized model that can predict failures in hot areas like Phoenix or even in extreme cold areas like Minneapolis. This research could have important implications for diagnosing and predicting electronic component failures at the vehicle level and could be widely applied to other components as well.

Acknowledgement

This study was supported by the National Research Foundation of Korea (Grant No. 2022R1A2C2011034). This work was supported by Korea Evaluation Institute of Industrial Technology (Grant No. 20018208).

References

Zhao, X., Fang, Y., Min, H., Wu, X., Wang, W., & Teixeira, R. (2023). Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective. *Expert Systems with Applications*, 121358.

Gültekin, Ö., Cinar, E., Özkan, K., & Yazıcı, A. (2022). Real-time fault detection and condition monitoring for

industrial autonomous transfer vehicles utilizing edge artificial intelligence. *Sensors*, 22(9), 3208.

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19-31.

Fang, Y., Min, H., Wu, X., Lei, X., Chen, S., Teixeira, R., & Zhao, X., (2023). Toward interpretability in fault diagnosis for autonomous vehicles: interpretation of sensor data anomalies. *IEEE Sensors Journal*, 23(5), 5014-5027.

Xu, Y., Sun, Y., Liu, X., & Zheng, Y. (2019). A digital-twin-assisted fault diagnosis using deep transfer learning. *IEEE Access*, 7, 19990-19999.

Li, X., Zhang, W., Ma, H., Luo, Z., & Li, X. (2020). Domain generalization in rotating machinery fault diagnostics using deep neural networks. *Neurocomputing*, 403, 409-420.

Li, Q., Chen, L., Kong, L., Wang, D., Xia, M., & Shen, C. (2023). Cross-domain augmentation diagnosis: An adversarial domain-augmented generalization method for fault diagnosis under unseen working conditions. Reliability Engineering & System Safety, 234, 109171.

Chang, F., Jafarzadeh, E., Del Gatto, J., Cran, G., & Sadjadi, H. (2023, October). Failure Mode Investigation to Enable LiDAR Health Monitoring for Automotive Application. *Annual Conference of the PHM Society*, Vol. 15, No. 1

Yuan, H., Cui, N., Li, C., Cui, Z., & Chang, L., (2023). Early stage internal short circuit fault diagnosis for lithium-ion batteries, *Journal of Energy Storage*, Vol. 57, 106196

Tang, L., Tian, H., Huang, H., Shi, S., & Ji, Q., (2022). A survey of mechanical fault diagnosis based on audio signal analysis, *Measurement*, Vol. 220, 113294

Weiss, K., Khoshgoftaar, T. M., & Wang, D., (2016). A survey of transfer learning, *Journal of Big data*, 3, 1-40.

BIOGRAPHIES

Sanghoon Lee received his B.S. and M.S. degrees in Automotive Engineering from Kookmin University, Seoul, South Korea, in 2008 and 2009, respectively. From 2009 to 2011, he worked as a research engineer at the Korea Institute of Science and Technology (KIST), Seoul, South Korea. Since 2011, he has worked at the Reliability and Certification Research Laboratory of the Korea Automotive Technology Institute (KATECH), Cheonan, South Korea, where he is currently a principal researcher. He is currently pursuing a Ph.D. in mechanical engineering from Yonsei University, Seoul, South Korea. His research interests include failure-based lifetime prediction and the health management of automotive systems.

Jaewook Lee received his B.S. degree in Mechanical Engineering from the Yonsei University, Seoul, South Korea in 2021. He is currently pursuing an Integrated M.S. and Ph.D. student in Mechanical Engineering at Yonsei University, Seoul, South Korea. His research interests are on the field of prognostics and health management (PHM), signal processing, deep learning-based data augmentation and knowledge transfer learning.

Jongsoo Lee received B.S. and M.S. in Mechanical Engineering at Yonsei University, Seoul, Korea in 1988 and 1990, respectively and Ph.D. in Mechanical Engineering at Rensselaer Polytechnic Institute, Troy, NY in 1996. After a research associate at Rensselaer Rotorcraft Technology Center, he has been a professor of Mechanical Engineering at Yonsei University since 1997. His research interests include multi-physics design optimization (MDO), reliability-based robust optimization, virtual product design and development, model-based system engineering (MBSE), prognostics and health management (PHM), and industrial artificial intelligence of data augmentation, transfer learning, domain adaptation, domain generalization and domain randomization with applications to structures, fatigue/durability, lifetime prediction, noise and vibration problems.

From Prediction to Prescription: Large Language Model Agent for Context-Aware Maintenance Decision Support

Haoxuan Deng[1,*], Bernadin Namoano[1], Bohao Zheng[1], Samir Khan[1], and John Ahmet Erkoyuncu[1]

[1]*School of Aerospace, Transportation and Manufacturing, Cranfield University, Bedford, MK43 0AL, UK*
{haoxuan.deng, bernadin.namoano, bohao.zheng, samir.s.khan, j.a.erkoyuncu}@cranfield.ac.uk

ABSTRACT

Predictive analytics with machine learning approaches has widely penetrated and shown great success in system health management over the decade. However, how to convert the prediction to an actionable plan for maintenance is still far from mature. This study investigates how to narrow the gap between predictive outcomes and prescriptive descriptions for system maintenance using an agentic approach based on the large language model (LLM). Additionally, with the retrieval-augmented generation (RAG) technique and tool usage capability, the LLM can be context-aware when making decisions in maintenance strategy proposals considering predictions from machine learning. In this way, the proposed method can push forward the boundary of current machine-learning methods from a predictor to an advisor for decision-making workload offload. For verification, a case study on linear actuator fault diagnosis is conducted with the GPT-4 model. The result demonstrates that the proposed method can perform fault detection without extra training or fine-tuning with comparable performance to baseline methods and deliver more informatic diagnosis analysis and suggestions. This research can shed light on the application of large language models in the construction of versatile and flexible artificial intelligence agents for maintenance tasks.

1. INTRODUCTION

Predictive analytics for product health management has attracted increasing attention from the industry with the rise of machine learning in the last decade. With the advent of advanced data processing and statistics methods, features and patterns of the system's running state can be captured from historical logs and sensor data. By doing this, potential system failure can be forecasted and allow people to outline the plan for maintenance or adjustment in advance of system deterioration. This can not only prolong the lifespan of the system but also lead to lower costs of periodic checking and overhauls in traditional preventive and reactive maintenance (Zonta et al., 2020).

Since the 2010s, deep learning that builds upon artificial neural networks (ANNs) played an essential role in predictive analytics and performed state-of-the-art results in many situations. Without tedious and complex feature engineering, deep learning can be effectively and efficiently applied to different data formats and draw relatively accurate predictions in an end-to-end way compared to other methods.

Even though the exciting breakthrough brought by deep learning for predictive maintenance, most of the research on this topic mainly focuses on boosting prediction metrics of the proposed methods such as precision or recalling rate, which provide limited information to the maintenance plan outlining (Roy et al., 2016). A higher prediction accuracy may indicate a more stable and reliable alarm in practical applications but does not necessarily suggest helpful decision support. It is practical and meaningful to know how to address an issue rather than merely anticipate it, especially in dealing with a complicated system containing numerous variables. Under this circumstance, predictions may only be treated as notifications and consequently ignored by human operators due to restricted proactive guidance. Thus, there is a strong call for extending machine learning beyond predictor to a more engaged advisor for action recommendation and insightful analysis (Matyas et al., 2017).

The mentioned main issue cannot be overcome by pure data-driven approaches based on statistics and algorithms since data collected from sensors only represent low-level signal patterns that are hard to analyze by human beings. Thus, it is difficult to form useful and helpful advice or guidance for decision-making (Sapna et al., 2019). To address this challenge, knowledge of contextual information is required to elaborate the prediction results into high-level representations such as natural language or graph-structured data so that human beings can view them straightforwardly. Hence, the industry calls for a more advanced agent system that can generate human-understandable descriptions for reviewing and validation based on detected faults.

The critical gap lies in the missing link between sources on the low-level data side and the high-level knowledge side.

In the last five years, there has been dramatic progress in the natural language process (NLP) because of the occurrence of large language models (LLMs). By pretraining a very deep neural network with billions of parameters on an extensive textual corpus, the LLMs can be multi-task learners with impressive performances on a wide range of tasks including article summarization, multilingual translation, and text generation. More importantly, recent research indicates the emergent capability of the LLMs for multi-step reasoning to accomplish more complex tasks without much human supervision and hardcore programming (Bommasani et al., 2021).

This exciting phenomenon indicates a solution to the mentioned challenge that the link can be regarded as a step-by-step transformation workflow starting from data to knowledge using LLMs with proper prompts. A basic idea is to allow LLMs to be aware of the fault in the system at first, and then parse relevant search queries related to the predicted fault for knowledge retrieval in the database. The obtained search results can then be combined and summarized as a document for action recommendation. In this way, the LLM is an information fusion unit to elaborate predictions with information from different databases for decision support.

According to this motivation, in this research, GPT-4, a popular large language model released by OpenAI (OpenAI et al., 2023), is applied to implement the above idea. The agent is built upon a fault classification model and external knowledge databases for the retrieval-augmented generation of the system maintenance support documentation. In addition, a use case on linear actuator fault diagnosis will be conducted for proof-of-concept verification. In summary, the contribution of the paper can be summarized as:

Develop an agent for linking prediction results with the knowledge base to generate descriptions for maintenance decision-support based on the large language model.

The remaining of the paper is organized in the following structure. In section two, some related work of this research will be presented for a preliminary introduction to the critical concept used in the proposed method. In section three, the system diagram and the framework will be illustrated in detail including the principles and workflow of the method. Then, there is a use case for linear actuator fault diagnosis will be conducted and experiments will be carried out to first show the effectiveness of methods for fault detection without extra training and fine-tuning. After that, another experiment will show how the LLMs can output a more context-related conclusion for a more satisfactory decision support delivery based on predictions.

2. RELATED WORK

2.1. Random Convolution Kernel Transformation (ROCKET) for Time Series Classification

For system state monitoring, multiple sensors will be installed on an asset to record a series of time-ordered data points during the system running. The collected time series will vary when the equipment works under different conditions, conversely, the time sequence data can represent in what situation the system is working and suggest what potential fault will probably be. To build the relationship between the time sequence with the corresponding system state, a classifier is the most effective way to implement, and this task is called Time Series Classification (TSC). It is one of the basic and essential time series mining that aims to assign unseen samples with labels in the training data by pattern exploitation. With TSC, a real-time collected time series can be categorized into states for a quick diagnosis. Therefore, the TSC is vital and commonly blind tightly to the industrial Internet of Things (IoT) for automatic system fault detection.

However, it is a challenge to apply conventional statistical or machine learning methods for the TSC. The main reason is the continuity property of the sequence of observable data points along time. Unlike textual data, which can be discretized by a set of sub-words (tokens) for processing, it is difficult to figure out the proper segmentation and transformation of the given time series for dimensionality reduction. This will cause an issue called the ‘curse of dimensionality’ and the model will be hard to recognize and capture discriminative features in the data for categorization.

There are fruitful results in TSC (Faouzi, n.d., 2022). Baseline methods such as K-nearest neighbors (KNN) classification with dynamic time warping (DTW), the bag-of-pattern method (BoP), and the remarkable ensemble classifier HIVE-COTE are proposed for this purpose, but they suffer from heavy computation and memory usage. Approaches based on deep learning such as recurrent neural network (RNN), InceptionTime (Fawaz et al., 2019), and relatively new Transformer-based models (Nie et al., 2022) are becoming popular. Although these methods boosted the accuracy and are able to generalize to different datasets compared to traditional machine learning, the model has to be trained to optimize parameters for inference which are either time-consuming or resource-intensive. Even worse, all the deep learning methods require retraining when samples are out of training data leading to a low extendibility in industrial applications.

To address these challenges, random convolution kernel transformation, short for ROCKET, was proposed to transform the time series into a vector representation using a random convolution kernel for classification (Dempster et al., 2019). Unlike conventional convolutional neural networks (CNN), parameters in the ROCKET are generated randomly and require no optimization or fine-tuning during the data transformation. Without an iterative learning process, the

ROCKET is efficient in computation and can adapt to different time sequences. Also, the ROCKET combined with traditional classification models such as the 1-NN classifier, support vector machine (SVM), and ridge classifier can achieve or even exceed state-of-the-art TSC algorithms with lightweight computation in an endurable timespan.

Therefore, considering the computational efficiency, extendibility, and performance, the ROCKET will be applied as the method for time series processing in this project. Using ROCKET as an encoder for the time series classification, the vectorization result will be processed and recognized by the LLMs to outline the prescription.

2.2. Retrieval-Augmented Generation (RAG)

Research on the application of LLMs in various workflow automation is conducted to unleash the power of LLM's human-like logical reasoning and inference capability. However, one of the main challenges comes in the *hallucination issue* of the LLMs. It means that the fake or incorrect information will be generated by LLMs. This can cause failure in task performance and may hurt the trustworthy between humans and the LLMs when they are in cooperation (Huang et al., 2023).

An effective approach for alleviating the hallucinating issue is to enable the LLMs to generate their responses based on some factual evidence from other existing sources such as the internet or knowledge databases. According to the retrieved information, the LLMs can follow the requirements and instructions given in the prompt to compile information that is rooted in ground truth and users' demands. It combines the searching techniques and the generation ability of LLMs to offer reliable and user-friendly information to people. This concept is defined as retrieval-augmented generation (RAG). The RAG has successfully been used in text, image, and multimodality searching and generation, but the application in the time series analysis on industrial sensor networks is yet fully explored. This piece of research is an initial exploration of applying the same idea for time series classification and allowing the LLM can generate the document based on the prediction to mitigate the hallucinations. In this research, the RAG will be the main methodology for retrieving historical time series samples to label the newly arrived data as a prediction outcome. The fault analysis can then be generated based on the retrieved result by the LLM.

2.3. Prompting Engineering, Chaining, and LLMs Agent

To obtain desirable outcomes from the LLMs, it is crucial to craft proper instructions for model commanding, and this concept is referred to as prompt engineering. Depending on the emergent capability, the LLMs can generate responses following descriptions in the prompts, and this is an effective way to alleviate the hallucination issue. Despite the usefulness, it is also cumbersome to tune the proper prompt to get satisfactory outcomes in a trial-and-error way.

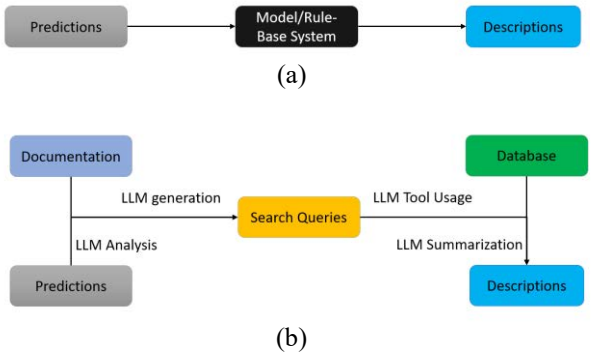


Figure 1. Different methods to convert predictions to descriptions (a). End-to-end generation (b). Multi-step transformation with the LLM

Instead of composing zero-shot prompts fully manually for LLMs to arrive at solutions immediately, (Wei et al., 2022) proposed chain-of-thought prompting that breaks down complex tasks into sequential sub-tasks and encourages the LLMs to figure out answers to each problem. By doing this, the LLMs can enhance their ability to successfully carry out intricate tasks such as math reasoning and arithmetic computations. In addition, (Yao et al., 2022) developed the ReAct prompting to enable LLMs to incorporate external tools usage and observation results obtained after tools leverage into their reasoning activity. The experiment indicated an apparent improvement in performance for LLMs on interactive text-based games and online shopping tasks as compared to traditional imitation learning or reinforcement learning approaches.

Furthermore, multiple prompts for different purposes can be serialized into a chain for workflow automation. Building on this advantage, the concept of the LLM agent, or AI agent, emerges to facilitate more functional applications of LLMs across diverse domains. The fundamental principle is to treat the LLMs as a connector or a controller among toolsets including databases, calculators, and web browsers to produce a series of actions based on their logical reasoning. In each step, LLMs can yield more reliable intermediary results and merge findings from prior stages to aggregate a more solid outcome in the final. Moreover, users can monitor the problem-solving process and understand the rationale provided by the agent, offering an opportunity for human intervention via a natural language interaction. Preliminary successful implementations of the LLMs agent in various fields are illustrated in (Xi et al., 2023).

3. Design of the Context-Awareness Agent

3.1. Initial Analysis

To compile a report of fault diagnosis with fault type, fault description, and potential recovery or maintenance strategies,

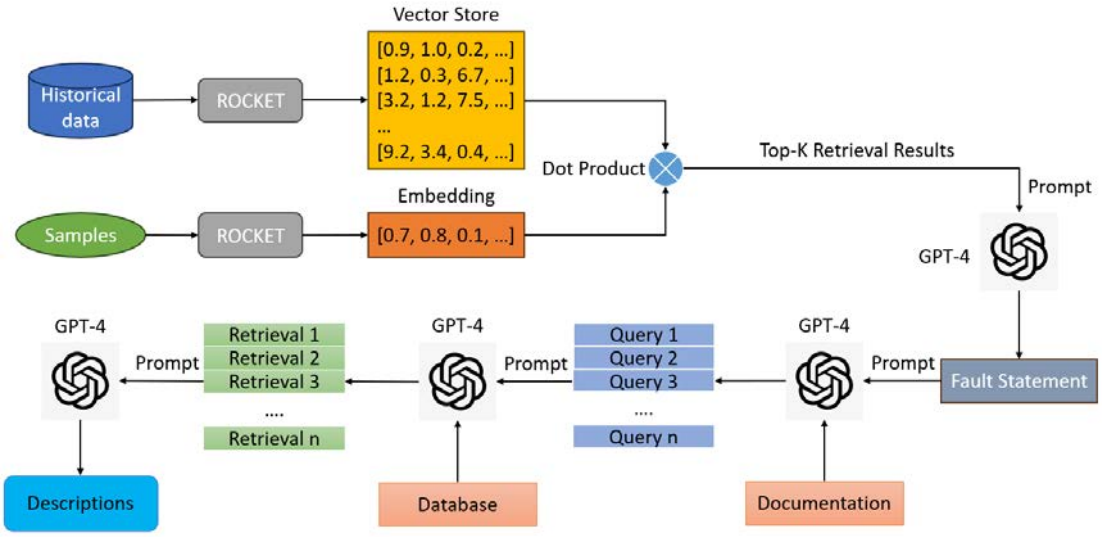


Figure 2. The diagram of the agent system

a direct way, as shown in Figure 1. (a), is to define a rule-based system that allows the prediction to go through and route to a corresponding solution. Or to craft a dataset of prediction-solution pairs to train a model for transformation. However, defining rules and collecting datasets manually are human-labor intensive and time-consuming. Also, the predefined rule-based system or model is hard to update and extend to unseen situations. From this perspective, end-to-end generation may be friendly in the development stage but may be challenging to maintain when the model or the system has been deployed in the production environment in industrial environments due to low adaptability to a dynamically evolving situation.

Another way to consider this problem is to divide the transformation into a series of steps with the LLM, external database, and tools (usually a bunch of calling Application Programming Interface), as shown in Figure 1. (b). After receiving the prediction from the data processing step, LLM will be asked to recognize the fault based on the prediction result and try to reason about what problem should be solved based on providing the contextual background described in the documentation. The generated search queries will then be thrown to a database by LLM's tools calling capabilities to retrieve relevant information on repairing suggestions. Finally, the LLM can be instructed to summarize all information into a report for human beings to review.

In this way, the workflow can be independent of a fixed set of rules and ensure contextual information is involved in the final description generation. In each step, human operators can track and offer comments or feedback to inject their expertise into the agent by providing prompting to get a more comprehensive result. The essential idea of the proposed method is *the multi-step generations based on factual evidence*, which is the core idea of the RAG. In the following section, the details of the agent system will be illustrated.

3.2. Overview of the Agent System

The diagram of the agent system is illustrated in Figure 2. A set of historical data will be transformed into a set of vectors and stored in memory for comparison. The method for sample vectorization is the ROCKET as introduced in the above section. In detail, several randomly generated convolution operators will slide the input time series to conduct dot product computation. In mathematics, according to the paper (Dempster et al., 2019), the outcome from implementing a kernel, w, with dilution, d, and bias b, on a specific set of time series X from position i in X_i is presented as follows:

$$X_i * w = \left(\sum_{j=0}^{l_{kernel}-1} X_{i+(j\times d)} \times w_j\right) + b \quad (1)$$

A feature map M will be obtained from the kernel computation, and two real values will be extracted as features for each kernel including the *maximum values* and the *proportion of positive values* (*ppv*) in the M by the following formula:

$$ppv(M) = \frac{1}{n}\sum_{0}^{n-1}[m_i > 0] \quad (2)$$

Where m_i is the numerical value in the feature map. Therefore, there will be two features produced per kernel operation, and for an effective time series representation, 10000 kernels are used to transform the data leading to 20000 features to represent each time series. The ROCKET algorithm is applied to all samples in the historical database which will be stored for retrieval. Since there is no parameter optimization and fine-tuning during the data processing, the computational efficiency can be extremely fast compared to

deep learning or other statistical methodologies. Therefore, the requirement on hardware configuration is much lower enabling a constant vectorization of new samples as experience accumulation.

When an unlabeled sample comes, it will also be transformed into a vector or said embedding and compute the Euclidean distance among all embedded samples in the vector store. After sorting based on distance, the most similar records will be considered as the target and the label will be assigned to the new data for classification result. For a common RAG implementation, in each retravel, the first five similar, or said the top-5 similar records will be extracted to promise a high hitting rate. In this agent system, top-5 retrieval is applied.

In the next step, the top-5 similar retrieval results will be fed to the LLM with a prompt to warp the prediction with contextual information including background introduction, technical details, and the system configuration to form a fault diagnosis statement in the following format:

Retrieval results: ['fault type1', 'fault type2', ..., 'fault type5']

Diagnosis results: ['fault type']

Inference evidence: [fault type1 with <score1>, ...]

Description of the Fault: This state indicates that…

In this compact diagnosis report, retrieval results will be shown, and the classification is presented as 'fault type'. In addition, inference evidence is the list of scores of the retrieval. In this case, the Euclidean distance is used for an interpretable purpose so that people can understand how the system gets the result. The more similarity between the unlabeled one and the records, the smaller the Euclidean distance will be. The description of the fault is summarized in the given document to explain to human operators clearly what is happening in the system with plain natural language.

In the next step, the brief statement is fed back to the LLM later and the fault type will be recognized for parsing the query for searching the database. For example, if the detected fault is 'spalling' on a ball-screw actuator given in the statement, the LLM can generate highly related several searching strings:

- How to recover spalling damage in ball-screw actuators?
- What are replacement options for ball-screw actuators with spalling damage?
- Replacement options for ball-screw actuators with spalling damage.
- Diagnosing spalling in linear actuators for effective maintenance.

These questions are then used for matching contents in a database, for instance, a general knowledge base e.g. Wikipedia, or a specific expert system with the tool usage capability of the LLM. All the obtained information will be summarized to direct maintenance suggestions and action recommendations in the final step.

4. USE CASE ON LINEAR ACTUATOR FAULT DIAGNOSIS

4.1. Experimental Setup

The time series data[1] is collected on a linear actuator system reported in the paper (Ruiz-Carcel & Starr, 2018). The detailed description including the mechanical components, structure, and parameters of configuration are all illustrated clearly in the article. This paper will not fully reintroduce the actuator system. The dataset acquired during the testing is the starting point of the introduction to the agent system application use case.

At first, the rig was operated under typical working conditions without any malfunctions to gather a substantial volume of data that represents the system's behavior under varying loads and motion patterns. Two distinct motion profiles were examined:

- Trapezoidal profile with a constant speed set point
- Sinusoidal profile with a smooth transition speed

In this paper, only data under the trapezoidal profile is considered for simplicity, the utilization of multiple profiles can be taken into account in future upgrades. The trapezoidal profile is tested for normal and faulty conditions under three distinct load scenarios: 20kgf, 40kgf, and -40kgf. The full motion sequence was repeated 5 times in one working situation under a load as one test. Each test will be conducted 10 times repetitively to generate a dataset with an adequate amount of observation in each case studied. a total of 50 samples, in every scenario analyzed. Furthermore, three distinct mechanical flaws in different degradation levels were intentionally introduced into various portions of the system to simulate modes typically experienced by these types of machines. The faults of the system in this dataset include:

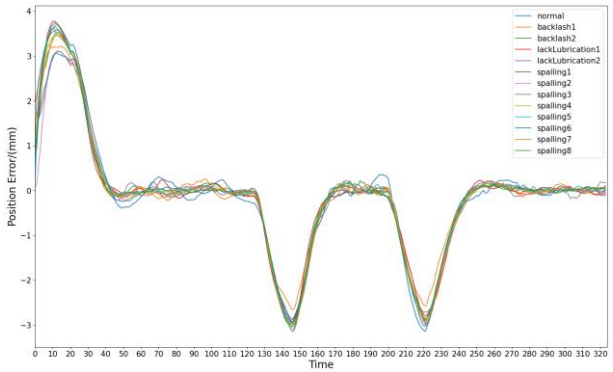
- Spalling from level 1 to level 8 (8 states)
- Lack of lubrication from level 1 to level 2 (2 states)
- Backlash from level 1 to level 2 (2 states)

In short, including the normal and all other flaw states, there are 13 different types and 650 samples in each load circumstance leading to a total of 1950 samples. For evaluation, 20% of all samples will be randomly selected to form a testing dataset, and the remaining samples will be used for constructing the vector store as introduced in section 3.2.

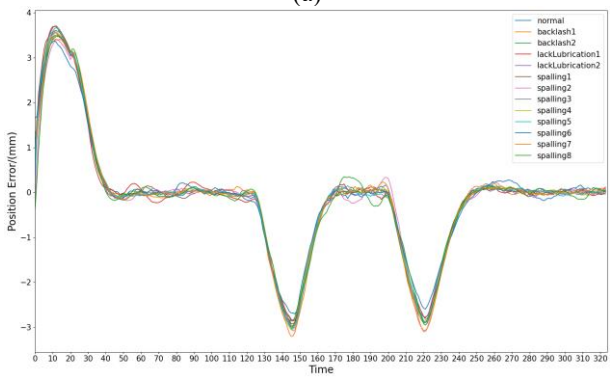
4.2. Implementation

For analysis, samples under all 13 flaws in each load can be visualized in Figure 3 and Figure 4 after using the moving average smoothing with the window size 20 and 15 respectively reported in the paper (Ruiz-Carcel & Starr,

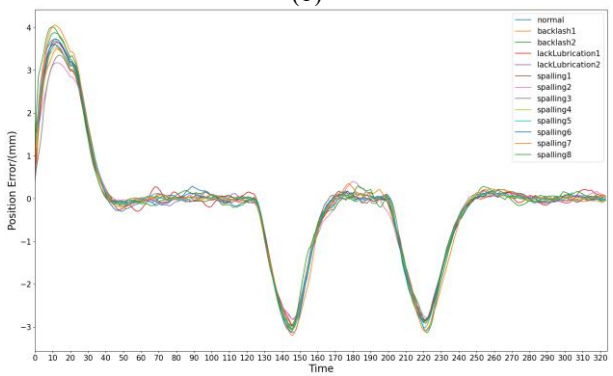
1.The dataset can be found on the link: https://cord.cranfield.ac.uk/articles/dataset/Data_set_for_Data-based_Detection_and_Diagnosis_of_Faults_in_Linear_Actuators_/5097649



(a)

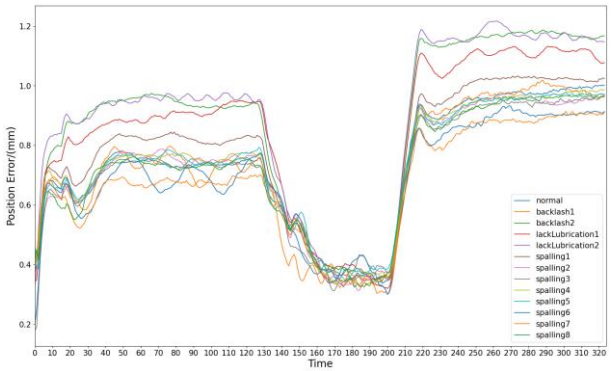


(b)

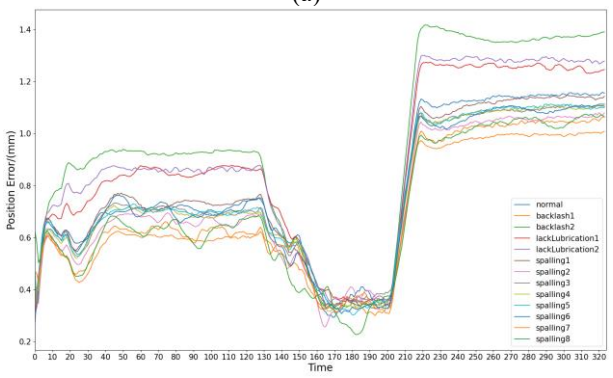


(c)

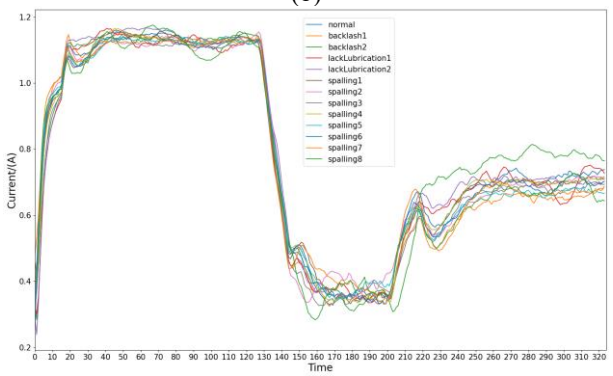
Figure 3. Position error (mm) signals in different fault states under three load conditions after smoothing. (a) 20kgf (b) 40kgf (c) -40kgf



(a)



(b)



(c)

Figure 4. Current (A) signals in different fault states under three load conditions after smoothing. (a) 20kgf (b) 40kgf (c) -40kgf

2018). As shown in Figure 3, no matter in which situation, the position error signals in each fault are distributed too close to be separated from others, while the pattern of current signals is more distinguishable. Therefore, the current signal is the univariant for time series processing in this use case.

Then the univariant time series will be transformed into a vector representation using the ROCKET. To further improve the computational efficiency, a trick from the variant of the ROCKET, namely MiniROCKET (Dempster et al., 2020), is applied. The main difference is that the kernel length usage is fixed to 9, instead of randomly selected from choices {7, 9, 11} in the original ROCKET. By doing this, it can make the result more deterministic. In this use case, the ROCKET is implemented with the Python package called Pyts (Faouzi & Janati, 2020).

Then the vector consisting of features computed from each kernel (20000 dimensions) will be stored together as a vector database for retrieval.

During the validation, prediction accuracy will be tested in three different loading conditions individually and the average value will also be computed. For the diagnosis report

Table 1. Fault classification accuracy under 20kgf, 40kgf, and -40kgf loading situations with different methods.

Method	20kg	40kg	-40kg	Average
InceptionTime (with 100 epochs)	83.8462%	86.0465%	79.2308%	83.0412%
ROCKET + top-1 retriever	63.0769%	75.1938%	46.1538%	61.4748%
ROCKET + ridge classifier	80.7692%	82.1705%	83.0769%	82.0055%
ROCKET + top-5 retriever	83.0769%	91.4729%	81.5385%	85.3628%

generation, given a loading, randomly select a sample from the testing dataset to review the fault statement and the relativity between the generated recommendations and the fault type.

4.3. Results

To summarize the results of the experiment, different methods for the classification based on the ROCKET features with Euclidean distance metric are listed in Table 1. The comparison between the proposed fault detection method is compared with the performance of the strong deep learning baseline method, the InceptionTime. There is an obvious gap between the top-1 precision based on the ROCKET vectorization and the baseline algorithm, only around 60% on average versus the InceptionTime, which is more than 80% over three loading circumstances. However, the computational time can be cut down dramatically by the ROCKET method. For the same historical dataset, for instance, using all 520 samples under 20kg, the result can be obtained with CPU (i7-12700H @3.30Hz) for about 25 seconds, while using the InceptionTime with 100 training epochs, the prediction drawn from scratch requires more than 60 seconds on GPU (Nvidia GeForce RTX 3060 Laptop GPU). Even though the deep learning model can quickly conclude after training, the parameters are fixed once the training is done. When out-of-distribution samples come, the model requires to be updated without forgetting previously obtained knowledge. Retraining or fine-tuning the model in this way is still an ongoing research topic. In contrast, using the ROCKET with the distance-based metric retriever, new samples can be encoded in nearly real-time to query existing vector databases to get the results. Thus, it shows the potential of on-the-fly data processing capability in industrial applications.

One way to further improve the classification accuracy is the incorporation of the ridge or logistic classifiers which can reach a better prediction outcome. As shown in Table 1, the ROCKET feature with the ridge classifier can even surpass InceptionTime in terms of prediction accuracy under the load of 40kg. In addition, an alternative approach is to use a top-5 retriever. By doing this, when one of the retrieved samples indicates the correct label, the prediction can be treated as correct. It can obviously boost the prediction accuracy (up to 85%) compared to any other top-1 classifiers, while with the cost of bringing the noise and uncertainty by considering more historical samples. Some types of faults may have highly similar patterns in the time series data, resulting in their simultaneous extraction as targets in the retrieval process. For instance, the ‘spalling2’ sequence can potentially retrieve ‘spalling1’ or ‘normal’ records from the vector database because their shapelets share great similarities. More importantly, this kind of uncertainty is usually unknown when deploying the system into the real production environment and delivering imprecise information to the users.

This problem can be addressed by refining the fault in a prompt in the brief fault diagnosis statement generation using the LLM. This is meaningful, as in real-life scenarios, system degradation occurs gradually and may not have a clear boundary or change in different malfunctions from a macro perspective. Therefore, two similar types of faults such as ‘spalling1’ and ‘spalling2’ may have nearly identical effects to the system and appear to have the same level of degradation. Thus, they can be regarded as the same fault type when proposing maintenance suggestions in practice. By reassigning these faults with a shared label in the prompt for the LLM, we can shrink the noise introduced by the top-5 retriever and take practical considerations into account while generating maintenance suggestions, which will make them more useful and accurate in real-life situations. This is the first step towards integrating realistic contextual information into the fault diagnosis process, making it more reliable and practical.

After getting the refined fault labels, the LLM (in this case, the GPT-4) can be instructed for multiple query generation, database retrieval, and the final summarization with handcrafting prompts step-by-step as illustrated in Figure 2 (The completed prompt can be found in the project repository listed in the Appendix). This can compile the final diagnostic report for converting the classification result to an actionable plan. Also, the Google Chrome web browser is selected as a general knowledge base used for query searching in this proof-of-concept verification. As a result, Figure 5 shows a part of the generated document indicating that the GPT-4 can give context-aware suggestions to a detected fault.

5. Discussions and Future Work

The use case successfully demonstrates the feasibility of applying the LLM to convert the single prediction to the

Retrieval results: ['spalling7_40_1_5', 'spalling7_40_9_3', 'spalling7_40_10_4', 'spalling7_40_7_3', 'spalling7_40_10_2']
Diagnosis results: Refined fault type1: Obvious fault in spalling
Inference evidence: [spalling7_40_1_5 with 3.1987850822921007, spalling7_40_9_3 with 3.772884271423856, spalling7_40_10_4 with 3.989529120390713, spalling7_40_7_3 with 4.239072661571159, spalling7_40_10_2 with 4.822921558144848]
Description of the Fault: This state indicates that the ball-screw mechanism within the actuator has significant surface damage. This level of spalling affects the actuator's smoothness and efficiency severely. The presence of such a fault suggests that the actuator's performance and longevity are compromised, necessitating immediate attention and likely repair or replacement to restore normal operation.

Searching for:
how to repair spalling damage in ball-screw actuators
best practices for preventing spalling in linear actuators.
replacement options for ball-screw actuators with spalling damage
diagnosing spalling in linear actuators for effective maintenance
cost-effective solutions for spalling damage in actuator mechanisms

Searching results:
Reconditioning the Actuator's Ball Screw: This involves repairing and/or replacing worn or damaged internal components of the actuator.
Replacing Cover Bands: When necessary, the cover bands should be replaced to ensure the proper functioning of the actuator.
Addressing Premature Flaking or Abnormal Damage: This may include looking for the cause of the problem, such as careless handling, excessive misalignment, or insufficient lubrication, and taking measures to prevent machine failure.
Checking for Damage: It is crucial to check for metal fragments that can indicate broken parts and could cause further damage. Also, inspect ball nut threads for damage and replace them if necessary.
Troubleshooting and Countermeasures: Check for specific damage conditions like the raceways of the screw shaft and ball nut and/or the surface of the ball peeling off like scales. Implement countermeasures based on identified causes, such as misalignment or lack of lubrication.
Seeking Professional Help: For severe damage or complex repairs, consulting with or hiring a professional service that specializes in actuator repair might be beneficial. Emergency repair teams can restore damaged ball screws and actuators when time is critical.
Observing Safety and Proper Procedures: Always follow safe handling practices and the manufacturer's instructions to minimize the risk of further damage during the repair process.
…

Repair and Reconditioning Strategies
To address immediate damage and restore functionality, the following steps are vital:
Ball Screw Reconditioning: Essential for repairing or replacing the internal parts of the actuator to improve functionality.
Cover Bands Replacement: Required when these components are damaged to maintain efficiency.
Addressing Underlying Causes: Investigating and rectifying root causes like misalignment or insufficient lubrication to prevent recurrence.
Damage Inspection: Identifying broken parts through inspection is crucial for preventing further damage.
Implementing Countermeasures: Specific damage conditions should be examined and addressed accordingly, emphasizing the correction of misalignment and proper lubrication.
Professional Consultation: Seeking expert advice is recommended for complex or severe damages.
Safe Handling and Procedure Adherence: Following manufacturer guidelines and safe practices is imperative throughout the repair process.
…

In summary, the effective management of spalling in balls crew actuators involves a balanced approach of timely and appropriate repairs, adherence to best practices for prevention, and the consideration of replacement when necessary. The combination of condition-based maintenance and leveraging cost-effective technological solutions plays a crucial role in enhancing the longevity and reliability of these actuators, ultimately ensuring their optimal performance in various applications.

Figure 5. A partial piece of an example document for the randomly selected ‘spalling7’ fault.

prescription of repairing and reconditioning strategies for maintenance decision support. Without any extra training or fine-tuning, and no requirement on manual feature engineering, dataset construction, and rule-based system definition, the LLM can automatically link different (public or private) knowledge sources to compile a reasonable solution after the fault diagnosis based on humans’ intention. Therefore, it improves the functionality of current machine learning as a more proactive and user-friendly production for industrial applications. In addition to the work presented in this paper, there are a few interesting directions that can be explored in the future.

From Univariant to Multivariant: In this study, only univariant time series (the current signal) is considered for the ROCKET feature construction, while the data related to position error has been ignored due to the similar shape patterns in the time domain. This raises the question of how to incorporate multiple time series patterns into the vector store section for fault diagnosis. The ROCKET can be extended to process multivariant time series, and it is a candidate update to the proposed agent system with this capability. In addition, all the introduced methods are in the time domain, how to integrate information from the frequency domain into the proposed framework is another question. By doing this, the time series can be analyzed from different points of view and construct more distinguishable features for retrieval with less amount of data.

From Suggestion to Automation: The report generated after the workflow is expected to offer assistance to people in maintenance planning. A further idea is to explore how to

connect the decision with the action to automate the entire maintenance process from fault detection and identification to system recovery and reconfiguration. A quick idea is to extend the sequentialization of prompts till to execution stage by incorporating external application programming interfaces (APIs) to directly link to actuators for the system maintenance. Recent relevant research is conducted for this purpose such as code-as-policy (Liang et al., 2023). In this way, it is exciting to develop an automated agent that can be self-awareness, self-decision, and self-action to the system health management without too much human intervention. However, it is also important to note that the verification and evaluation from the human side are critical to ensure the final action satisfies all practical and aesthetic requirements in production environment. How the agent can learn from human feedback to further align their performance with our expectations and values is a critical consideration for this direction.

From Ad-hoc Prompting to Long-Term Memory: The prompts used in this paper are yet fully automatically generated. Handcrafting is still needed during the prompting process. For every generation, the GPT-4 should reload all information from scratch and provide suggestions merely limited to the information written in the prompt. Hence, the agent cannot view and refer to any of previous diagnostic reports to improve its performance and keep accumulating experience for future analysis. Some studies show that if an agent can learn from the contents generated by itself, after self-learning on these contents, the performance may improve and even exceed the human level. AlphaGo, for instance, can self-play with enormous virtual games by itself and eventually defeat top-ranked human players (Silver et al., 2017). A further question is whether this similar idea can be applied to the agent. If the prompting and previous diagnosis reports can be stored in long-term memory and retrieved for new situations by the agent itself. It is possible to let the agent itself to prompt itself automatically with lower human supervision. This can cut down the requirement of human knowledge and computation time. Also, it can increase the likelihood of producing more optimal solutions that people have yet to conceive.

From the Given Knowledge Base to Self-Exploration: It is noticeable that automation is built upon a human-defined logic written in prompts. Thus, the basis of automation still relies much on human labor and insight. More importantly, the knowledge base for agent retrieval is also created and mainly maintained by human beings, thus, heavily restricting the potential of machine knowledge discovery. It is then followed by a question of how to allow the machine to acquire knowledge with the self-exploring capability to discover new methods to alleviate human bias and errors in maintenance tasks. Furthermore, it is fascinating to investigate how to enable the agent to contribute to the existing knowledge with human beings together for knowledge acquisition in the system health management domain.

6. CONCLUSION

This paper presents an LLMs agent-based method for elaborating predictions from machine learning to actionable strategy descriptions for maintenance decision support. A use case of linear actuator fault diagnosis is studied with an agent built upon ROCKET time series representation, the concept of RAG, and the prompts chaining technique. By prompting engineering, the LLM agent can recognize the fault and parse highly relevant queries to the database using a search tool, (in this case, the Google Chrome web browser), and summarize the retrieval results to report to human operators. The study demonstrates the possibility of constructing autonomous agents for proactive decision assistance without much human supervision and training and shows how current LLMs can be integrated into the industrial workflow.

ACKNOWLEDGEMENT

This research was supported by the Center for Digital Engineering and Manufacturing at Cranfield University (UK)

APPENDIX

The code for this project can be found on the link:
https://github.com/BlueAsuka/Rocket-RAG

REFERENCES

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. https://arxiv.org/abs/2108.07258v3

Dempster, A., Petitjean, F., & Webb, G. I. (2019). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, *34*(5), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z

Dempster, A., Schmidt, D. F., & Webb, G. I. (2020). MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 248–257. https://doi.org/10.1145/3447548.3467231

Faouzi, J. (n.d.). *Time Series Classification: A review of Algorithms and Implementations*. Retrieved March 23, 2024, from https://inria.hal.science/hal-03558165

Faouzi, J., & Janati, H. (2020). pyts: A Python Package for Time Series Classification. *Journal of Machine Learning Research*, *21*(46), 1–6. http://jmlr.org/papers/v21/19-763.html

Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2019). InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962. https://doi.org/10.1007/s10618-020-00710-y

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. https://arxiv.org/abs/2311.05232v1

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., & Zeng, A. (2023). Code as Policies: Language Model Programs for Embodied Control. *Proceedings - IEEE International Conference on Robotics and Automation*, *2023-May*, 9493–9500. https://doi.org/10.1109/ICRA48891.2023.10160591

Matyas, K., Nemeth, T., Kovacs, K., & Glawar, R. (2017). A procedural approach for realizing prescriptive maintenance planning in manufacturing industries. *CIRP Annals*, *66*(1), 461–464. https://doi.org/10.1016/J.CIRP.2017.04.007

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). *A Time Series is Worth 64 Words: Long-term Forecasting with Transformers*. https://arxiv.org/abs/2211.14730v2

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. https://arxiv.org/abs/2303.08774v6

Roy, R., Stark, R., Tracht, K., Takata, S., & Mori, M. (2016). Continuous maintenance and the future – Foundations and technological challenges. *CIRP Annals*, *65*(2), 667–688. https://doi.org/10.1016/J.CIRP.2016.06.006

Ruiz-Carcel, C., & Starr, A. (2018). Data-Based Detection and Diagnosis of Faults in Linear Actuators. *IEEE Transactions on Instrumentation and Measurement*, *67*(9), 2035–2047. https://doi.org/10.1109/TIM.2018.2814067

Sapna, R., Monikarani, H. G., & Mishra, S. (2019). Linked data through the lens of machine learning: An Enterprise view. *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*. https://doi.org/10.1109/ICECCT.2019.8869283

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature 2017 550:7676*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2201.11903v6

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., … Gui, T. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey*. https://github.com/WooooDyy/LLM-Agent-Paper-List.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models*. https://arxiv.org/abs/2210.03629v3

Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*, *150*, 106889. https://doi.org/10.1016/J.CIE.2020.106889

Fully Automated Diagnostics of Induction Motor Drives in Offshore Wind Turbine Pitch Systems using Extended Park Vector Transform and Convolutional Neural Network

Manuel S Mathew[1], Surya Teja Kandukuri[2], Christian W Omlin[3]

[1,3]*University of Agder, Jon Lilletuns vei 9, 4879 Grimstad, Norway*
manuel.s.mathew@uia.no
christian.omlin@uia.no

[2]*Norwegian Research Centre, Energy & Technology Department, Tullins Gate 2, 0166 Oslo, Norway*
suka@norceresearch.no

ABSTRACT

Due to their location and related complexities, the offshore wind farms (OWF) have higher downtimes and operation and maintenance (O&M) costs compared to their onshore counterparts. Condition monitoring could help in bringing down the O&M costs of OWFs. The pitch system is one of the components most prone to failure. This paper details an approach for enhanced diagnosis of the electric pitch systems especially focusing on the induction motor drives (IMD) in wind turbines. The proposed method uses an extended Park vector approach (EPVA) in conjunction with a convolutional neural network (CNN) to accurately classify the condition of an IMD and localize the faults. The method is validated on data collected from a laboratory setup. The advantage of the proposed approach is that the condition of the IMD can accurately be classified, and faults localized in operating conditions with varying load and frequency without any additional information on the instantaneous operating speed, frequency, or load on the motor drives. This results in a non-invasive diagnostic approach incurring least additional expenses to implement.

1. INTRODUCTION

Offshore wind farms (OWF) have significant potential to contribute towards global energy sustainability. However, they face unique operational challenges, mainly because of their remote locations and harsh marine environments in which they operate. To put this in perspective, while onshore wind farms are attaining a 95% to 97% availability for modern systems Pfaffel, Faulstich and Rohrig (2017), the availability of OWFs is relatively lower and highly variable. The data from earlier offshore wind farms suggest an availability of 67% to 85% (Feng, Tavner, & Long, 2010) with more latest estimates of 80% to 84% (Cevasco, Koukoura, & Kolios, 2021). The limited weather windows for performing necessary maintenance leads to longer downtimes, which explains the gap in operational availability between onshore and offshore wind farms. Furthermore, the operational and maintenance (O&M) costs constitute a significant proportion of the lifetime costs associated with OWFs with estimates ranging from roughly 23% on the lower end (Ren, Verma, Li, Teuwen, & Jiang, 2021) to 30% at the higher end of the spectrum (Hammond, & Cooperman, 2022). This represents a stark contrast to onshore wind farms, where the lifetime O&M costs typically account for approximately 5% (Ren et al., 2021). Thus, implementing condition-based maintenance (CBM) strategies, and hence condition monitoring (CM) become vital in reducing the costs associated with O&M and helps in reducing the downtimes in maintenance activities.

The pitch system of wind turbines is among the components most prone to failures and one that contributes significantly to a non-trivial amount of downtime. The results from ReliaWind project (Wilkinson et al., 2010) indicate the pitch system was responsible for nearly 15% of failures per turbine per year and close to 20% of total downtime hours per year across different manufacturers in their database. A more recent study (Walgern, Fischer, Hentschel, & Kolios, 2023) suggests a pitch system failure rate of 0.54 (hydraulic) and 0.56 (electrical) per turbine per year. Pitch systems are also found to be the most critical subcomponent in the premature failure period (Santelo, De Oliveira, Maciel, & De A. Monteiro, 2022). This makes the pitch system an ideal candidate for enabling CM systems, particularly in OWFs because of the additional costs and downtimes associated with their reactive maintenance.

Although there have been some efforts to develop CM solutions for wind turbine pitch systems, the extent of these attempts has not been commensurate with their critical impact on downtime and failure rates. Cho, Gao and Moan (2016) developed a Kalman filter based method for diagnosing pitch sensor and actuator faults in floating wind turbines based on the NREL 5MW wind turbine model. However, the focus here was not on the incipient fault detection. Several machine learning-based techniques for fault diagnostics in pitch systems are also found in the literature. supervisory control and data acquisition (SCADA) data has been used to detect anomalies in the pitch system in tandem with Isolation Forest based anomaly detection models (Mckinnon, Carroll, Mcdonald, Koukoura, & Plumley, 2021). In this work, the authors developed different models of varying amounts of training data to detect anomalous patterns. Further, they experimented with varying lengths of post-processing window to see how it affects their model. Their results show that their method could notice signs of turbine failure 12 to 18 months ahead. Park, Kim, Dinh and Park (2022) used neural networks to find abnormal operations in the pitch system of a wind turbine. The authors define the abnormal operation for the pitch system using the deviation in the blade pitch angle, where a deviation of more than 4.95° in blade pitch angle was considered abnormal. Wei, Qian and Zareipour (2019) developed a condition monitoring and fault detection system for the wind turbine pitch system using optimized relevance vector machine regression. Their work leverages the SCADA data to detect faults in the pitch system particularly focusing on encoder failures, pitch controller failures, electric motor failures, and slip ring failures. Similarly Chen, Matthews and Tavner (2013) used SCADA data to develop an a-priori knowledge-based ANFIS (APK-ANFIS) model to detect faults in the wind turbine pitch system. The authors identified four critical characteristics features (CF) of pitch faults after analyzing data in the developmental stage of a fault and that immediately after the maintenance has been carried out. These CFs have been used to develop the corresponding APK-ANFIS models, the results from which where aggregated to detect the fault in the pitch systems. Most of the studies in the literature reviewed are effective in detecting faults in the pitch systems, however, they fall short of delineating the fault diagnosis to a subcomponent level.

A subcomponent level diagnosis of faults is essential as it contributes to efficient planning and implementation of maintenance activities, especially in OWFs, where precise planning is paramount. While SCADA data can be effectively used for preliminary fault diagnosis, it is less effective in the subcomponent level fault diagnosis. In this paper, we focus on fault diagnostics for the induction motor drive (IMD) of an electrical pitch system. Subcomponent level fault diagnosis for pitch motor drives using current signature analysis have been previously addressed by the authors (Kandukuri, Karimi, & Robbersmyr, 2016; Kandukuri, Senanayaka, Huynh, Karimi, & Robbersmyr, 2017) , and also proposed a two-stage fault classification scheme based on support vector machine (SVM), for large-scale deployment in OWFs (Kandukuri, Senanyaka, & Robbersmyr, 2019).

The issue with classical current signature-based methods in fault detection is that there is an assumption of steady state operations in terms of speed and load. The wind turbine pitch systems on the other hand are operated intermittently and are exposed to varying speed and load profiles. This means that, either regions of steady state operations must be carefully detected for data acquisition or advanced signal processing techniques are to be employed (Benbouzid, M El Hachemi, 2000; Bhole, & Ghodke, 2021; Liu, & Bazzi, 2017).

Thus, in this paper, a novel solution is proposed by calculating the extended Park vector (EPV) current from the three-phase motor line currents and then extracting the time-frequency representation using Short-Term Fourier Transform (STFT). The three-phase motor line currents for this purpose are observed at varying operating conditions: speed, and load. For detecting the condition of the IMD, these representations are subsequently converted into spectrograms, which are then used to train a convolutional neural network (CNN) for classification of the IMD's condition. CNNs have earlier been reported focusing on diagnostics of gearboxes (Amin, Bibo, Panyam, & Tallapragada, 2023; Gecgel, Ekwaro-Osire, Gulbulak, & Morais, 2021; Jiang, Han, & Xu, 2020), bearings (Choudhary, Mian, & Fatima, 2021; Lu et al., 2023; Ruan, Wang, Yan, & Gühmann, 2023; Wang, Mao, & Li, 2021; Yuan, Lian, Kang, Chen, & Zhai, 2020), and IMDs (Junior et al., 2022; Khanjani, & Ezoji, 2021; Kumar, & Hati, 2022; Lee, Pack, & Lee, 2019; Skowron, Orlowska-Kowalska, Wolkiewicz, & Kowalski, 2020). However, most of these works depend on vibration sensors, or are specific to one type of fault. Further, most of them assume constant supply frequency and load. Skowron et al. (2020) warrants a special mention as they use motor line currents to detect faults in an IMD. They normalize these currents to create a vector that is then reshaped to form an RGB matrix corresponding to each of the three phases. While they were able to detect and differentiate between various kinds of stator faults including insipient faults, their work still deals with one kind of fault within the induction motor.

Thus, what differentiates this work from other works are as follows:

1. Fault diagnostics of IMDs operating under varying speed, frequency, and load conditions without needing any additional data on these parameters.
2. Beyond identifying a single type of fault, the proposed approach is capable of fault localization using extended Park vector approach (EPVA) in conjunction with a CNN classifier.

3. The proposed approach, through EPVA, negates the need for additional sensors to be deployed. This makes it an economically viable option for wind turbines without vibration sensors, especially those that are nearing their end of designed life. Because while vibration sensor-based diagnostics are more widely applicable, they are expensive (Trajin, Regnier, & Faucher, 2010) compared to motor current signature analysis (MCSA).
4. Since continuous monitoring of the WT pitch system is not required in this method, intermittent snapshots of the three-phase currents are sufficient for reliable diagnosis. Thus, reducing the data transmission load from each turbine.
5. The proposed algorithm need not be implemented at each turbine, the proposed approach can contribute towards farm-level health management.

While EPVA and similar MCSA methods have earlier been used for IM diagnosis and prognosis (Erik Leandro, Levy Ely De Lacerda De, Jonas Guedes Borges Da, Germano, & Luiz Eduardo Borges Da, 2012), to the best of authors' knowledge this is the first time the diagnostics of the induction motor has been fully automated while using spectrogram of the EPVA in conjunction with a deep learning based classifier. The rest of the paper is organized as follows. In section 2, the induction motor faults under consideration and the reason for selecting these are discussed. This is followed by a brief explanation of theories of EPVA and CNNs in section 3. Section 4 details the laboratory setup used to collect the necessary data and Section 5 discusses the methodology of research and details about the CNN based classifier. Results from the classification scheme are discussed in section 6. The paper is concluded, and possible future directions are briefly highlighted in section 7.

2. IMD FAULTS CONSIDERED

Despite being robust, induction motors are not immune to failures. Stator faults and bearing faults are among the most reported components contributing to the total failures in an IMD (Benbouzid, M., 1999; Benbouzid, M. E. H., & Kliman, 2003; Singh, & Ahmed Saleh Al Kazzaz, 2003; Thorsen, & Dalva, 1995) as shown in Figure 1. Nearly half of the total failures are the result of stator faults, making it one of the most important types of faults to be detected. This is followed by bearing faults which account for almost one-third of the total failures. Compared to those, a menial 10% of the failures are accounted for by rotor faults. These faults occur generally because of drive-generated harmonics, poor ventilation at low-speed operation, and abrupt load variations.

Thus, in this study, we have considered the two components contributing the most to the total IMD failures: stator fault and bearing fault. A future study may be done including rotor bar faults such as different severities of broken rotor bar (BRB) faults.

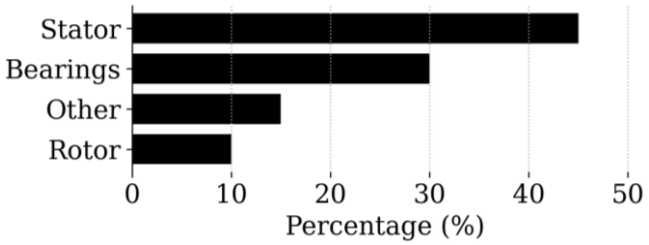


Figure 1. Distribution of IMD faults

3. THEORY

3.1. Extended Park Vector (EPV) Analysis

The theory behind using MCSA for IMDs revolves around the concept that an induction motor, while operating in healthy state, is symmetrical across the three phases. A fault in the motor disrupts this symmetry causing a periodically recurring asymmetry in the motor's operational characteristics. This periodic recurrence manifests as a particular frequency in the current known as "fault frequencies" or "signature frequencies."

These fault frequencies arise due to the interaction between the motor's electrical and mechanical components influenced by the fault. For example, a stator fault, such as insulation failure, due to short circuits between the stator windings, or phase imbalance, causes an asymmetric distribution of electromagnetic fields (EMF) within the motor. This asymmetry causes variations in the magnetic effect on the rotor resulting in irregular rotor motion. The interaction between the EMF and the rotor's motion produces specific frequency components in the motor's currents called "stator fault frequencies". These stator fault frequencies are then reflected in the MCSA as harmonics of the fundamental frequency, or appearance of specific sidebands around the fundamental frequency and its sidebands. In the case of a bearing fault, which can occur because of physical damage, wear and tear, inadequate lubrication, or external factors, leads to mechanical vibrations which modulate the EMF within the motor affecting the air gap flux density. This introduces specific frequencies in the motor's current signature called "bearing fault frequencies". The specific frequency of the bearing fault depends on several factors like bearing design, motor speed, and the nature of the fault. These signature frequencies are then used to diagnose the faults in MCSA.

EPV analysis builds upon the foundational principles of MCSA and has been used for a while now in diagnosing motor electrical faults (Cardoso, Cruz, & Fonseca, 1997). The direct (i_d) and quadrature (i_q) axis currents are initially calculated as a function of three phase motor currents (i_a, i_b, i_c) as follows:

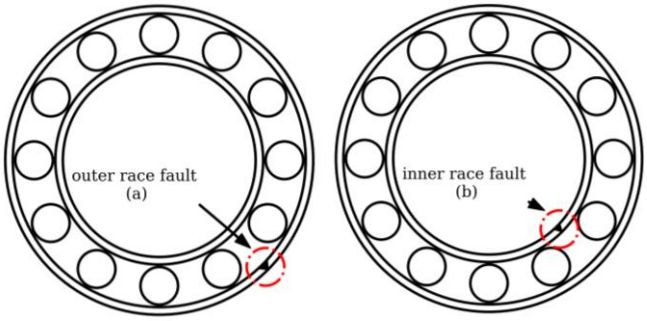


Figure 2. Schematic diagram of (a) an outer race fault, and (b) an inner race fault in a rolling element bearing.

$$i_d = \left(\sqrt{2}/\sqrt{3}\right) i_a - \left(1/\sqrt{6}\right) i_b - \left(1/\sqrt{6}\right) i_c \quad (1)$$

$$i_q = \left(1/\sqrt{2}\right) i_b - \left(1/\sqrt{2}\right) i_c \quad (2)$$

The extended Park vector (i_p) is then calculated as follows:

$$i_p = \left|(i_d + j i_q)\right| \quad (3)$$

Where, j is the imaginary unit defined as $j^2 = -1$.

When a stator turn fault (STF) occurs due to shorting between the phase windings, the three phase currents become imbalanced and the ideal values for direct and quadrature axis currents mentioned in Cardoso et al. (1997) doesn't hold anymore. The result is that the stator turn fault can be identified only using the spectral component at twice the supply frequency, f_s (Sahraoui, Zouzou, Ghoggal, & Guedidi, 2010) in the spectrum of i_p:

$$f_{STF} = 2f_s \quad (4)$$

Figure 2 shows the schematics of an outer race fault (a), and an inner race fault (b) in a rolling element bearing. A bearing fault, as discussed earlier, causes spectral components at different frequencies as determined by bearing design, motor speed, and the nature of the fault (e.g., faults on the inner race, outer race, ball spin, or cage defects). This results in three additional spectral components in the spectrum of i_p along with the fundamental component of the power supply (Zarei, & Poshtan, 2009) as:

$$f_{BRG} \in \{f_v, 2f_v, |2f_v - f_s|\} \quad (5)$$

For an outer race fault, as shown in Figure 2 (a), the characteristic vibration frequency, f_v, can be estimated using the following equation (Zarei, & Poshtan, 2009):

$$f_v \approx 0.4 N_b f_r \quad (6)$$

where N_b is the number of rolling elements in the bearing, and f_r is the shaft rotational frequency.

Even though these frequencies (f_{STF}, *and* f_{BRG}) can reliably be used for fault diagnosis in short time windows of constant operation, this fails in the case of a variable load and frequency because of the changes in the shaft rotational frequency, f_r, and supply frequency, f_s.

Characterizing the time-frequency response of the extended Park vector, i_p is critical in EPVA. The short-term Fourier transformations (STFT) is used to decompose the Park vector into its time-frequency components, which offers an in-depth view of how these frequency components evolve over time. This level of detail is more suitable for diagnosing the faults within motors operating under non-stationary conditions.

STFT is a special case of Fourier transforms where the Fourier transform is applied in series to smaller slices of the signal. The assumption here is that for a shorter time window, the original non-stationary signal becomes stationary. The STFT of a non-stationary signal $y(t)$ can be estimated by discretizing the continuous-time signal to a discrete-time signal, $y(n)$, where n is the discrete time indices. Then the discrete STFT is calculated for the discrete-time signal as:

$$Y(\omega, b) = \sum y(n) w(n - b) e^{-j\omega n} \quad (7)$$

where $w(.)$ is the windowing function and b is the window-shifting time constant. The calculation of STFT is done using a fixed-size window, which means that if the window is longer, frequency resolution is better at the expense of time resolution and vice versa for shorter windows. Thus, deciding a window length for the STFT operation is crucial in accurately extracting the desired time-frequency information (Oppenheim, 1999) and thereby diagnosing the motor condition.

3.2. Convolutional Neural Networks (CNN)

Though there has been some precedents in computer vision research inspired by natural vision, CNNs developed by Lecun et al. (1989) were instrumental in the development of computer vision at scale. CNNs are similar to artificial neural networks or "vanilla neural networks" in that they are made up of neurons. However, CNNs generally consists of three types of layers, namely convolutional layer, pooling layer, fully connected layers, and an output layer.

The convolutional layers are used to learn a feature representation from the inputs provided to create feature maps. In this layer, a learned kernel convolves with the input producing a feature map, the result of which is then passed on to an elementwise non-linear activation function to get the activation maps. Each element of the feature map is connected to a local subset of neurons in the previous layer or the input. The feature map element at (i, j) in the k^{th} feature map of the l^{th} layer can be calculated as:

$$z^l_{i,j,k} = \boldsymbol{w}^l_k \boldsymbol{x}^l_{i,j} + b^l_k \quad (8)$$

where $\boldsymbol{w}^l_k$ and b^l_k are the weight vector and bias term of the k^{th} filter of the l^{th} layer, respectively. $\boldsymbol{x}^l_{i,j}$ is the local subset of

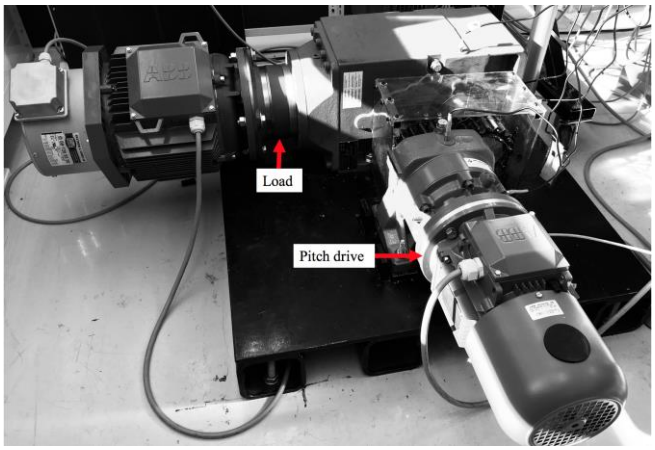


Figure 3. Laboratory setup for motor diagnostics

input to the convolutional layer centered at (i, j). An activation function such as rectified linear unit (ReLU) later introduces non-linearities helping the network to learn non-linear features.

The pooling layer, often placed between two convolutional layers, introduces shift-invariance to the feature maps. This is achieved by reducing the resolution of the feature maps. Usually, average pooling and max pooling layers are used depending on the task at hand. Higher-level feature representations are extracted eventually by stacking several convolutional and pooling layers.

One or more fully-connected layers usually succeed in CNNs aiming to achieve high level reasoning (Simonyan, & Zisserman, 2014). The last layer of CNNs is an output layer, which uses a task appropriate activation function such as sigmoid function for classification or ReLU for regression problems.

4. Laboratory Setup

Figure 3 shows the laboratory setup built to study the common faults in the pitch motor drives and planetary gear boxes of a wind turbine. A 1.1 kW, three-phase induction motor served as the test motor. Another 2.2 kW three-phase induction motor was used to supply the loads in the setup through a bevel-planetary-helical gearbox. Both the motors were driven by commercial field-oriented control (FOC) drives. Further details of the setup can be found in Table 1. The selection of the current sensor for this setup has been influenced by the desire for a common industrial sensor which is economical for installation on multiple units. Further, the overall frequency content of the signal has been tested over ideal power source and showed excellent signal-to-noise ratio. The speed and torque references for both test and load motors are provided to their respective FOC drives through a PC.

Table 1. Details of the test setup

Test Motor	
IM Rated Power	1.1 kW
IM Rated Speed	1420 rpm
IM Rated Torque	7.2 Nm
Current Sensor	
Model	LEM LTS-6NP
Primary nominal RMS current, I_{PN}	6 A
Accuracy @ I_{PN}, 25° C	±0.2
Data Acquisition	NI USB DAQ
Acquisition rate	15 kHz

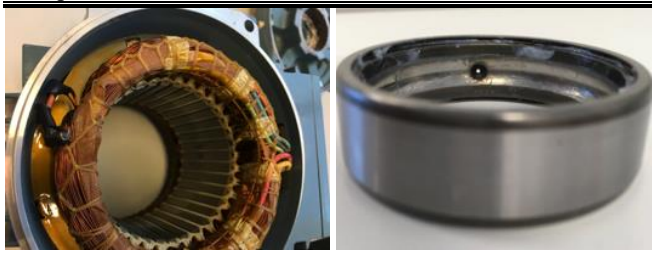


Figure 4. Seeded motor faults: stator turns fault (left), bearing fault (right).

The STF, and BRG faults were artificially seeded as shown in Figure 4. The STF was seeded by shorting 10% of a phase winding, while BRG was seeded as an outer race fault with a diameter of a ≈ 2mm through hole.

5. Methodology

Initially, the test motor was run in healthy condition, across a range of speeds varying between 850 rpm and 1420 rpm. At the same time, the loads were varied between no-load and full-load conditions at random to simulate the random loading on the wind turbine's pitch system. The speed interval was used after consulting the motor loadability curves to ensure that the motor reaches neither an overload condition nor stall condition because of the random loading. The randomness of operational conditions was ensured using different random number generators with seeds refreshed every thirty second interval. The three-phase currents were recorded as snapshots of 30 seconds each. Similarly, records were made for the motor operating in faulty conditions as mentioned in Section 4. A total of nearly a thousand minutes of data was collected from the test setup for this purpose.

Further, the Park vector, i_p and its STFT has been calculated for each of the recorded snapshots using equations (1), (2), (3), and (7). Examples of the STFT results from each of the three conditions: healthy, STF, and BRG is shown in Figure 5–7. Classical signal processing methods to detect faults from the STFT of the Park vector may fail here, however, from the figure, it can be noted that there is an increase in frequency content around 100 Hz in the STF condition, which is around $2f_s$ (Figure 6). A similar increase in frequency content can

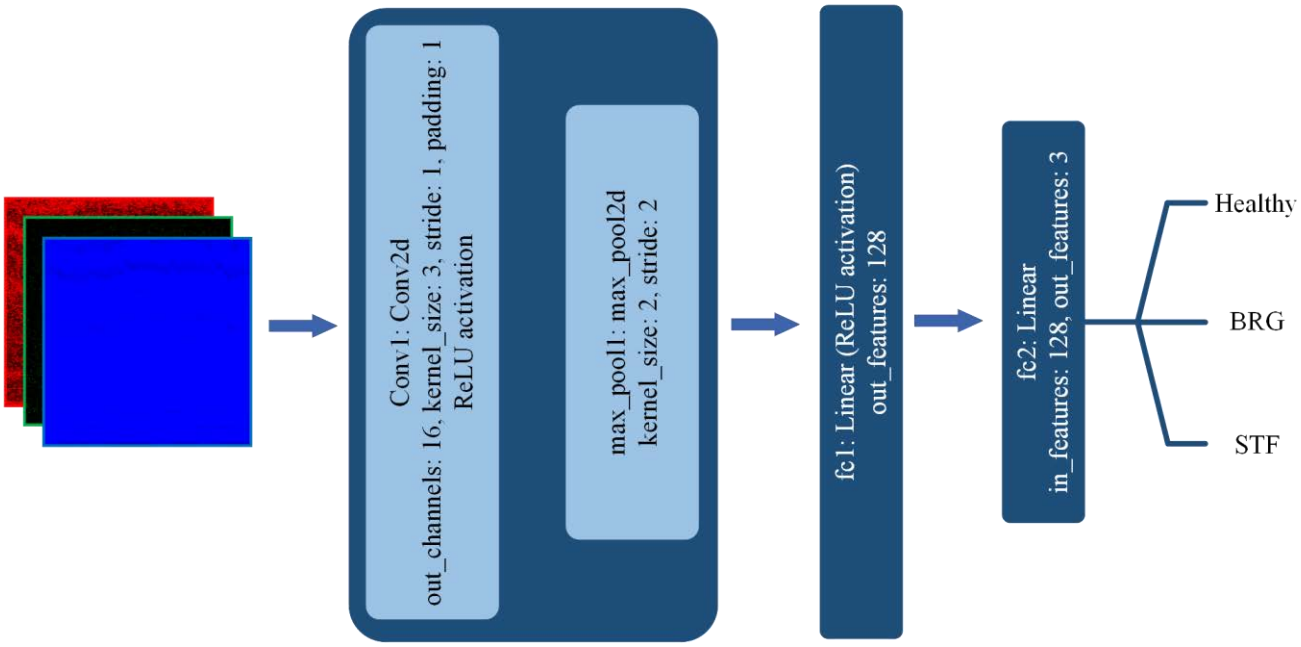


Figure 8. CNN architecture

also be observed in the case of BRG faults at around 500 – 600 Hz (Figure 7), which is distinct from the healthy case (Figure 5).

Around 2100 STFT images containing the time-frequency information associated with the operation of the motor in the three different states mentioned in the previous paragraph were then used to develop a CNN model. The entire dataset of images was split at random and 70% of the data was used in training while 15% each was used for validation and testing purposes. The STFT spectrum of the current, i_p, is obtained as an RGB image, which was then resized to 360 x 360. The architecture of the CNN that was developed for fault classification is shown in the Figure 8. The architecture consists of a convolutional layer followed by a max pooling layer, the output from which is flattened and forwarded to a fully connected layer. This fully connected layer learns high-level features from the flattened inputs. The final layer serves as the classifier, which takes the output from the fully connected layer to classify the image into one of the three conditions previously mentioned.

The CNN was trained on a system equipped with an Intel Xeon processor, NVIDIA Tesla V100-SXM3-32GB GPU on Python 3.10 using PyTorch 2.2. The model was trained over hundred epochs with a batch size of 256 and a learning rate of 10^{-3}. The learning rate was arrived at after narrowing down the value by using a learning decay scheduler. Early stopping and L2 regularization were employed to mitigate the possibility of model overfitting to the training dataset. The early stopping algorithm checks for any improvements in the validation loss and stops training if no improvement is

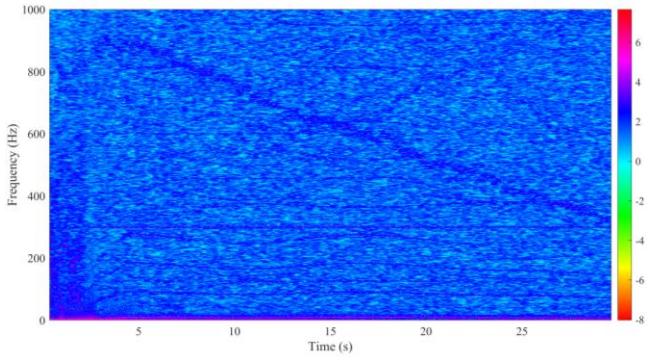


Figure 5. STFT of a Park vector, i_p, of motor working in healthy condition.

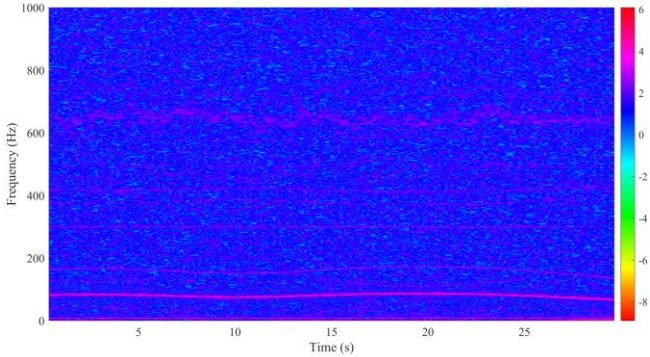


Figure 6. STFT of a Park vector, i_p, of motor with STF fault.

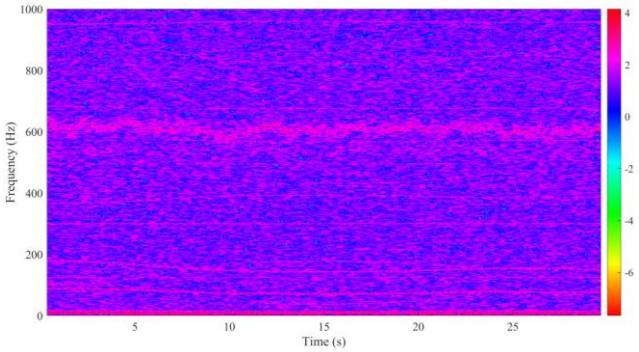


Figure 7. STFT of a Park vector, i_p, of motor with BRG fault.

observed for 25 epochs. Since this is a multi-class classification problem, the cross entropy loss was used as the loss function and the Adam optimizer (Kingma, & Ba, 2014) was selected for optimizing the loss function. At the end of each epoch the model was validated with the validation dataset and results recorded, which is detailed in the following section.

6. Results and Discussions

Figure 9 and Figure 10 illustrates the feature maps generated after the convolutional layer and max pooling layers of the trained CNN model in BRG and STF fault conditions, respectively. It is clear from the figures that the convolutional layer followed by the max pooling layer effectively identifies the specific locations within the spectrum associated with each fault condition.

After training, the model was tested on a previously unseen test dataset. Inference on GPU takes slightly higher than 17 seconds and that on CPU takes approximately 36 seconds to classify the 320 snapshots. Table 2 shows the confusion matrix of the model's performance on this dataset. The STF fault was the most accurately classified among the three

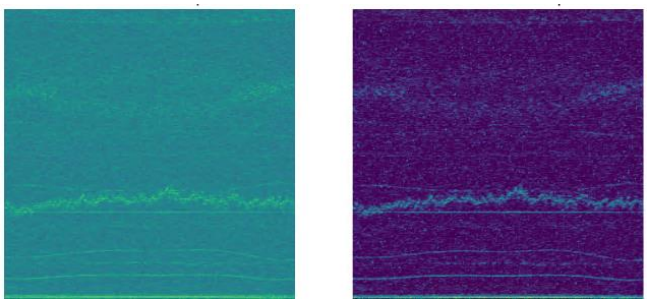


Figure 9. One of the feature maps after convolutional layer (left) and max pool layer (right) in BRG fault condition.

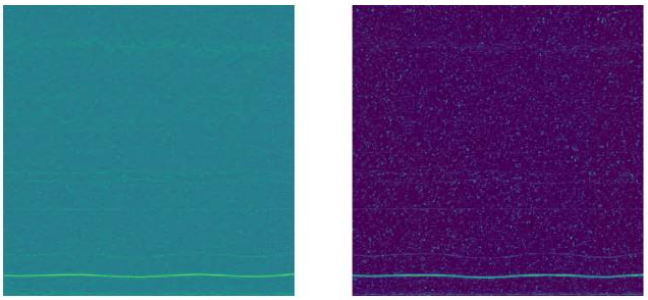


Figure 10. One of the feature maps after convolutional layer (left) and max pool layer (right) in STF fault condition.

Table 2. Confusion matrix of the model performance on test data.

	Healthy	STF	BRG	Percentage
Healthy	94	0	3	96.9%
STF	0	109	0	100%
BRG	4	0	110	96.5%

conditions with 100% of the cases being correctly identified as such. On the other hand, the model makes some mistakes while classifying the healthy and BRG fault conditions.

Table 3 shows the performance of the developed model in classifying the three motor conditions. The model has an overall accuracy of 97.8%. Similarly high values of precision, recall, and F1-score are observed when the model encounters the test dataset. Such high numbers might raise the suspicion of overfitting or data leakage, which is the case where the test dataset was inadvertently used in training the model. Early stopping and L2 Regularization help in preventing overfitting of the model on the training dataset while the entire data pipeline has been verified manually to ensure that there is no data leakage. Thus, the performance as shown in Table 3 highlights that the model has effectively learned the underlying pattern and can classify the three motor conditions effectively. Further, these high values in performance also indicate that the model is sufficiently complex to match the problem's complexity.

Table 3. Performance of the CNN model on test dataset.

Metric	Healthy	STF	BRG	Overall
Accuracy	-	-	-	97.8%
Precision	95.9%	100%	97.3%	97.7%
Recall	96.9%	100%	96.5%	97.8%
F1-Score	96.4%	100%	96.9%	97.78%

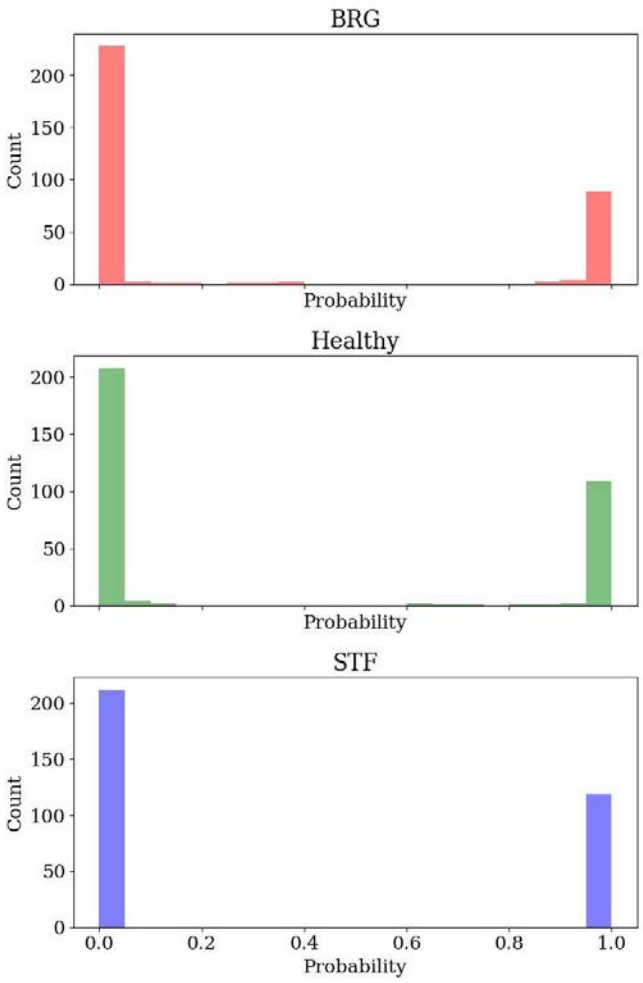


Figure 11. Histogram of predicted probabilities for each motor condition in test dataset.

To further explore the inference pattern, a histogram of the predicted probabilities was drawn across different classes as shown in Figure 11. Histogram of predicted probabilities illustrates the confidence of the model in classification to different classes. The pronounced skewness towards the extremities of the probabilities indicates a high level of certainty in classification. This is especially true for the STF fault where the model has a hundred percent confidence if a given snapshot is indicative of an STF fault (probability = 1.0) or not (probability = 0.0). On the other hand, the model is slightly less confident in identifying the healthy condition or the bearing fault as evidenced by relatively higher variance in the model's predicted probabilities.

7. Conclusion

In this paper, we have successfully developed and demonstrated a method for enhanced diagnostics for IMDs used in wind turbine pitch system. In addition to detecting the operational state of the motor (healthy/faulty), this approach also helps in localizing the fault to the stator or the bearing of the IMD. The presented approach uses the extended Park vector approach (EPVA) and short-term Fourier transforms (STFT) to extract the time-frequency information from the three-phase induction motor currents taken as 30-second snapshots. These snapshots are then classified using a convolutional neural network. The high accuracy in classification of the conditions indicates that the model can accurately diagnose the state of the IMD in question.

The advantage of the proposed method is that CNN requires only snapshots of 30 seconds each to determine the state of operation. Other than the sampling frequency, no additional information is required about the loading conditions or the frequency of operation, making it a suitable candidate for farm-level implementation. Further, as continuous monitoring is not required in this approach, it is ideal for WT pitch systems that operate intermittently. Since only 30-second snapshots of motor currents are the requirement, the data transferred from the WTs will be minimal.

The results presented here are tested on different motors of the same type, further validation of the methodology over different test motors could help strengthen the study in future. Additionally, at present two different fault conditions have been studied, more faults like broken rotor bar (BRB) faults or different stages of the STF could also be included in future works.

While the CNN classifier can be used to determine the motor condition accurately, the bottleneck in this methodology is the STFT calculations, which are computationally intensive. Thus, as a next step, the authors intend to test different ML paradigms to bypass the STFT calculations and directly detect these conditions from the Park vector current, i_p.

Acknowledgements

This research work has been funded by Analytics for asset Integrity Management of Windfarms (AIMWind), under grant no. 312486, from Research Council of Norway (RCN).

AIMWind is collaborative research from University of Agder, Norwegian Research Center (NORCE), and TU Delft, with DNV and Origo Solutions as advisory partners.

References

Amin, A., Bibo, A., Panyam, M., & Tallapragada, P. (2023). Vibration based fault diagnostics in a wind turbine planetary gearbox using machine learning. *Wind Engineering*, *47*(1), 175-189. https://doi.org/10.1177/0309524x221123968

Benbouzid, M. (1999). Bibliography on induction motors faults detection and diagnosis. *IEEE Transactions on Energy Conversion*, *14*(4), 1065-1074.

Benbouzid, M. E. H. (2000). A review of induction motors signature analysis as a medium for faults detection.

IEEE Transactions on Industrial Electronics, *47*(5), 984-993.

Benbouzid, M. E. H., & Kliman, G. B. (2003). What stator current processing-based technique to use for induction motor rotor faults diagnosis? *IEEE Transactions on Energy Conversion*, *18*(2), 238-244. https://doi.org/10.1109/TEC.2003.811741

Bhole, N., & Ghodke, S. (2021). Motor current signature analysis for fault detection of induction machine–a review. 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE),

Cardoso, A. J. M., Cruz, S. M. A., & Fonseca, D. S. B. (1997, 18-21 May 1997). Inter-turn stator winding fault diagnosis in three-phase induction motors, by park's vector approach. 1997 IEEE International Electric Machines and Drives Conference Record,

Cevasco, D., Koukoura, S., & Kolios, A. J. (2021). Reliability, availability, maintainability data review for the identification of trends in offshore wind energy applications. *Renewable and Sustainable Energy Reviews*, *136*, 110414. https://doi.org/10.1016/j.rser.2020.110414

Chen, B., Matthews, P. C., & Tavner, P. J. (2013). Wind turbine pitch faults prognosis using a-priori knowledge-based anfis. *Expert Systems with Applications*, *40*(17), 6863-6876.

Cho, S., Gao, Z., & Moan, T. (2016). Model-based fault detection of blade pitch system in floating wind turbines. Journal of Physics: Conference Series,

Choudhary, A., Mian, T., & Fatima, S. (2021). Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images. *Measurement*, *176*, 109196. https://doi.org/10.1016/j.measurement.2021.109196

Erik Leandro, B., Levy Ely de Lacerda de, O., Jonas Guedes Borges da, S., Germano, L.-T., & Luiz Eduardo Borges da, S. (2012). Predictive maintenance by electrical signature analysis to induction motors. In Prof. Rui Esteves, A. (Ed.), *Induction motors* (pp. Ch. 20). IntechOpen. https://doi.org/10.5772/48045

Feng, Y., Tavner, P. J., & Long, H. (2010). Early experiences with uk round 1 offshore wind farms. *Proceedings of the Institution of Civil Engineers - Energy*, *163*(4), 167-181. https://doi.org/10.1680/ener.2010.163.4.167

Gecgel, O., Ekwaro-Osire, S., Gulbulak, U., & Morais, T. S. (2021). Deep convolutional neural network framework for diagnostics of planetary gearboxes under dynamic loading with feature-level data fusion. *Journal of Vibration and Acoustics*, *144*(3). https://doi.org/10.1115/1.4052364

Hammond, R., & Cooperman, A. (2022). *Windfarm operations and maintenance cost-benefit analysis tool (wombat)* (NREL/TP-5000-83712). https://www.nrel.gov/docs/fy23osti/83712.pdf

Jiang, Z., Han, Q., & Xu, X. (2020). Fault diagnosis of planetary gearbox based on motor current signal analysis. *Shock and Vibration*, *2020*, 8854776. https://doi.org/10.1155/2020/8854776

Junior, R. F. R., Areias, I. A. d. S., Campos, M. M., Teixeira, C. E., da Silva, L. E. B., & Gomes, G. F. (2022). Fault detection and diagnosis in electric motors using 1d convolutional neural networks with multi-channel vibration signals. *Measurement*, *190*, 110759. https://doi.org/10.1016/j.measurement.2022.110759

Kandukuri, S. T., Karimi, H. R., & Robbersmyr, K. G. (2016). Fault diagnostics for electrically operated pitch systems in offshore wind turbines. Journal of Physics: Conference Series,

Kandukuri, S. T., Senanayaka, J. S. L., Huynh, V. K., Karimi, H. R., & Robbersmyr, K. G. (2017). Current signature based fault diagnosis of field-oriented and direct torque–controlled induction motor drives. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, *231*(10), 849-866.

Kandukuri, S. T., Senanyaka, J. S. L., & Robbersmyr, K. G. (2019). A two-stage fault detection and classification scheme for electrical pitch drives in offshore wind farms using support vector machine. *IEEE Transactions on Industry Applications*, *55*(5), 5109-5118.

Khanjani, M., & Ezoji, M. (2021). Electrical fault detection in three-phase induction motor using deep network-based features of thermograms. *Measurement*, *173*, 108622. https://doi.org/10.1016/j.measurement.2020.108622

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, P., & Hati, A. S. (2022). Dilated convolutional neural network based model for bearing faults and broken rotor bar detection in squirrel cage induction motors. *Expert Systems with Applications*, *191*, 116290. https://doi.org/https://doi.org/10.1016/j.eswa.2021.116290

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

Lee, J.-H., Pack, J.-H., & Lee, I.-S. (2019). Fault diagnosis of induction motor using convolutional neural network. *Applied Sciences*, *9*(15), 2950. https://www.mdpi.com/2076-3417/9/15/2950

Liu, Y., & Bazzi, A. M. (2017). A review and comparison of fault detection and diagnosis methods for squirrel-

cage induction motors: State of the art. *ISA transactions*, *70*, 400-409.

Lu, H., Pavan Nemani, V., Barzegar, V., Allen, C., Hu, C., Laflamme, S., Sarkar, S., & Zimmerman, A. T. (2023). A physics-informed feature weighting method for bearing fault diagnostics. *Mechanical Systems and Signal Processing*, *191*, 110171. https://doi.org/10.1016/j.ymssp.2023.110171

McKinnon, C., Carroll, J., McDonald, A., Koukoura, S., & Plumley, C. (2021). Investigation of isolation forest for wind turbine pitch system condition monitoring using scada data [Article]. *Energies*, *14*(20), 20, Article 6601. https://doi.org/10.3390/en14206601

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Park, J., Kim, C., Dinh, M. C., & Park, M. (2022). Design of a condition monitoring system for wind turbines [Article]. *Energies*, *15*(2), 16, Article 464. https://doi.org/10.3390/en15020464

Pfaffel, S., Faulstich, S., & Rohrig, K. (2017). Performance and reliability of wind turbines: A review. *Energies*, *10*(11), 1904. https://doi.org/10.3390/en10111904

Ren, Z., Verma, A. S., Li, Y., Teuwen, J. J. E., & Jiang, Z. (2021). Offshore wind turbine operations and maintenance: A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, *144*, 110886. https://doi.org/10.1016/j.rser.2021.110886

Ruan, D., Wang, J., Yan, J., & Gühmann, C. (2023). Cnn parameter design based on fault signal analysis and its application in bearing fault diagnosis. *Advanced Engineering Informatics*, *55*, 101877. https://doi.org/https://doi.org/10.1016/j.aei.2023.101877

Sahraoui, M., Zouzou, S. E., Ghoggal, A., & Guedidi, S. (2010, 6-8 Sept. 2010). A new method to detect inter-turn short-circuit in induction motors. The XIX International Conference on Electrical Machines - ICEM 2010,

Santelo, T. N., de Oliveira, C. M. R., Maciel, C. D., & de A. Monteiro, J. R. B. (2022). Wind turbine failures review and trends. *Journal of Control, Automation and Electrical Systems*, *33*(2), 505-521. https://doi.org/10.1007/s40313-021-00789-8

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, G. K., & Ahmed Saleh Al Kazzaz, S. a. (2003). Induction machine drive condition monitoring and diagnostic research—a survey. *Electric Power Systems Research*, *64*(2), 145-158. https://doi.org/https://doi.org/10.1016/S0378-7796(02)00172-4

Skowron, M., Orlowska-Kowalska, T., Wolkiewicz, M., & Kowalski, C. T. (2020). Convolutional neural network-based stator current data-driven incipient stator fault diagnosis of inverter-fed induction motor. *Energies*, *13*(6), 1475. https://www.mdpi.com/1996-1073/13/6/1475

Thorsen, O. V., & Dalva, M. (1995). A survey of faults on induction motors in offshore oil industry, petrochemical industry, gas terminals, and oil refineries. *IEEE Transactions on Industry Applications*, *31*(5), 1186-1196. https://doi.org/10.1109/28.464536

Trajin, B., Regnier, J., & Faucher, J. (2010). Comparison between vibration and stator current analysis for the detection of bearing faults in asynchronous drives. *IET electric power applications*, *4*(2), 90-100.

Walgern, J., Fischer, K., Hentschel, P., & Kolios, A. (2023). Reliability of electrical and hydraulic pitch systems in wind turbines based on field-data analysis. *Energy Reports*, *9*, 3273-3281. https://doi.org/10.1016/j.egyr.2023.02.007

Wang, X., Mao, D., & Li, X. (2021). Bearing fault diagnosis based on vibro-acoustic data fusion and 1d-cnn network. *Measurement*, *173*, 108518. https://doi.org/10.1016/j.measurement.2020.108518

Wei, L., Qian, Z., & Zareipour, H. (2019). Wind turbine pitch system condition monitoring and fault detection based on optimized relevance vector machine regression. *IEEE Transactions on Sustainable Energy*, *11*(4), 2326-2336.

Wilkinson, M., Hendriks, B., Spinato, F., Harman, K., Gomez, E., Bulacio, H., Roca, J., Tavner, P., Feng, Y., & Long, H. (2010). *Methodology and results of the reliawind reliability field study* European Wind Energy Conference , EWEC 2010, Warsaw, Poland. https://eprints.whiterose.ac.uk/83343/

Yuan, L., Lian, D., Kang, X., Chen, Y., & Zhai, K. (2020). Rolling bearing fault diagnosis based on convolutional neural network and support vector machine. *IEEE Access*, *8*, 137395-137406. https://doi.org/10.1109/ACCESS.2020.3012053

Zarei, J., & Poshtan, J. (2009). An advanced park's vectors approach for bearing fault detection. *Tribology International*, *42*(2), 213-219. https://doi.org/10.1016/j.triboint.2008.06.002

BIOGRAPHIES

Manuel S. Mathew is a PhD Research Fellow at the Information and Communication Technology department at the University of Agder, Norway. His interest is in the application of artificial intelligence in renewable energy systems particularly focusing on prognostics for wind farms. He completed his master's degree in Renewable Energy in 2021 from the University of Agder. In addition, he also holds a master's degree in systems engineering by research from the University of Brunei Darussalam. He did his bachelor's degree in electrical and electronics engineering from the Mahatma Gandhi University, India.

Surya Teja Kandukuri is a Senior Scientist at NORCE. He holds a part-time position as a researcher at University of Agder, Grimstad, Norway. He obtained his PhD in condition monitoring from the University of Agder in 2018. He has over 12 years of experience in industrial research within aerospace, energy, marine and oil & gas sectors, developing condition monitoring solutions for high-value assets. He received his master's degree in systems and control engineering from TU Delft, The Netherlands, in 2006 and bachelor's in electrical engineering from Nagarjuna University in India in 2003.

Christian W. Omlin has been a professor of Artificial Intelligence at the University of Agder since 2018. He has previously taught at the University of South Africa, University of the Witwatersrand, Middle East Technical University, University of the South Pacific, University of the Western Cape, and Stellenbosch University. His expertise is in deep learning with a focus on applications ranging from safety to security, industrial monitoring, renewable energy, banking, sign language translation, healthcare, bio conservation, and astronomy. He is particularly interested in the balance between the desire for autonomy using AI technologies and the necessity for accountability through AI imperatives such as explainability, privacy, security, ethics, and artificial morality for society's ultimate trust in and acceptance of AI. He received his Ph.D. from Rensselaer Polytechnic Institute and his MEng from the Swiss Federal Institute of Technology, Zurich, in 1995 and 1987, respectively.

Graph Neural Networks for Electric and Hydraulic Data Fusion to Enhance Short-term Forecasting of Pumped-storage Hydroelectricity

Raffael Theiler[1], Olga Fink[2]

[1,2] *EPFL, Lausanne, Vaud, 1015, Switzerland*
raffael.theiler@epfl.ch
olga.fink@epfl.ch

ABSTRACT

Pumped-storage hydropower plants (PSH) actively participate in grid power-frequency control and therefore often operate under dynamic conditions, which results in rapidly varying system states. Predicting these dynamically changing states is essential for comprehending the underlying sensor and machine conditions. This understanding aids in detecting anomalies and faults, ensuring the reliable operation of the connected power grid, and in identifying faulty and miscalibrated sensors. PSH are complex, highly interconnected systems encompassing electrical and hydraulic subsystems, each characterized by their respective underlying networks that can individually be represented as graph. To take advantage of this relational inductive bias, *graph neural networks* (GNNs) have been separately applied to state forecasting tasks in the individual subsystems, but without considering their interdependencies. In PSH, however, these subsystems depend on the same control input, making their operations highly interdependent and interconnected. Consequently, hydraulic and electrical sensor data should be fused across PSH subsystems to improve state forecasting accuracy. This approach has not been explored in GNN literature yet because many available PSH graphs are limited to their respective subsystem boundaries, which makes the method unsuitable to be applied directly. In this work, we introduce the application of *spectral-temporal graph neural networks*, which leverage self-attention mechanisms to concurrently capture and learn meaningful subsystem interdependencies and the dynamic patterns observed in electric and hydraulic sensors. Our method effectively fuses data from the PSH's subsystems by operating on a unified, system-wide graph, learned directly from the data, This approach leads to demonstrably improved state forecasting performance and enhanced generalizability.

1. INTRODUCTION

In power grids, pumped-storage hydropower plants (PSH) are well-established for large-scale energy storage due to their efficiency, scalablilty and flexibility. In this role, these plants dynamically respond to potentially large fluctuations in grid demand. In the transition towards smart grids, PSH sensor data is collected via *wide area measurement systems (WAMS)* (Pagnier & Chertkov, 2021a) and is stored in centralized *energy management systems (EMS)*. By processing the aggregated WAMS data, modern EMS provide crucial functionalities for PSH operators such as load forecasting, real-time monitoring, distribution and demand-side management, and various decision support tools aimed at increasing efficiency and sustainability. To enhance system reliability, EMS implement anomaly and sensor fault detection based on short-term forecasting and state estimation, playing a pivotal role in preventing failures that could lead to widespread power grid outages and significant economic losses. However, the dynamic operation of the PSH and the vast amounts of data transmitted by the WAMS significantly complicate the task. In the PSH environment, conventional state estimation is often ineffective because the computation can take several minutes (Li, Pandey, Hooi, Faloutsos, & Pileggi, 2022). This delay leads to a rapid divergence between the most recent and the previously used system state for the estimation, resulting in numerous false-positives when applied to anomaly detection. Consequently, it becomes challenging to maintain a comprehensive overview of the system's health and performance. As a solution, deep-learning-based short-term state forecasting has recently been applied, which offers significantly faster processing times and holds the potential to benefit from the additional data increasingly collected at a high sampling rate (Kundacina, Cosovic, & Vukobratovic, 2022).

Developing deep-learning-based state forecasting for PSH is particularly challenging. These challenges stem from the necessity to accurately represent two distinct physical domains within the PSH: the hydraulic and electrical systems. Al-

though these domains are mechanically interconnected by electromagnetic generators, they are traditionally modeled independently in mechanical engineering. This division mainly stems from the distinct dynamics governing each subsystem. It is, therefore, challenging to model hydraulic and electrical domains simultaneously. Nonetheless, considering the direct causal relationship between the systems – wherein kinetic energy is transformed to electric energy – we hypothesize that fusing data from both subsystems, which operate under a unified control input, can significantly enhance state forecasting.

To address the challenge of fusing electric and hydraulic data, we posit that both subsystems of the PSH consist of extensive networks, which are coarsely monitored with sensors that can be represented in the non-Euclidean graph domain. By operating on this more effective graph representation, which can capture biases given by the PSH system architecture and homophily biases, addressing the phenomenon that sensor measurements tend to be connected with "similar" or "alike" others (Ma, Liu, Shah, & Tang, 2023), *graph neural networks* (GNNs) have recently gained significant attention. When a graph is available, GNNs have been effectively applied in key applications to (hydro) power plants (Liao, Bak-Jensen, Radhakrishna Pillai, Wang, & Wang, 2022) and in the broader power grid environment. However, these methods depend on the availability of apriori graphs, derived from PSH's electrical and hydraulic network diagrams. Therefore, their applicability is limited by the fact that, although the underlying network structure of both electric and hydraulic subsystems of a PSH can be modeled as a graph, network diagrams for PSH exist typically only separately for each subsystem. Consequently, most graph-based methods are confined within the boundaries of their respective systems. To overcome this limitation, we propose learning a PSH sensor graph from latent dependencies in the data. While it has previously been demonstrated that graph structures can be efficiently learned from data, this approach remains unexplored in the context of hydropower plants. In light of this, inspired by (Cao, Li, Ma, & Tomizuka, 2021), we propose using *spectral-temporal graph neural networks* (STGNN) to learn a latent correlation graph structure across the entire PSH asset for the fusion of electric and hydraulic data, leveraging the self-attention mechanism.

Compared to numerical simulation, our data-driven GNN-based methodology is computationally inexpensive and does not require expert knowledge while maintaining interpretability, due to the accessibility of the learned graph. Our proposed approach can be easily transferred to different PSH assets without any calibration. To the best of our knowledge, there is no other work on data-driven electric and hydraulic data fusion for PSH using graph neural networks.

To summarize, in this work, we introduce the application of attention-based graph neural networks to effectively learn intra- and interdependencies between the subsystems' sensors to enhance the short-term state forecast in the pumped storage power plant (PSH) environment. We tackle several challenges when applying state forecasting to PSH:

- In line with the dynamic operation of the PSH, the dynamic behavior of sensors adds complexity to the forecasting. We propose a *spectral-temporal graph neural network* (STGNN) that effectively captures these patterns by incorporating the PSH' underlying structural and homophily biases, such as load patterns that are reflected across sensor measurement sites.
- State forecasting in the PSH environment depends strongly on environmental parameters, such as temperature, daily load profiles, and power grid customer-related factors, which cannot be modeled in numerical simulations (Lin, Wu, & Boulet, 2021). In contrast, our STGNN is able to learn these factors from data.
- The PSH is a spatially distributed complex system that spans across the hydraulic and electric domains. We propose a graph learning module that learns a unified graph representation across the hydraulic and electrical subsystems from latent dependencies in the data.
- We assess the performance of our method on a multivariate PSH dataset containing 58 signals, showcasing the dynamic operation of the asset.

The reminder of this paper is organized as follows: Sec. 2 reviews relevant literature that focuses on graph-based deep learning and data fusion. Sec. 3 introduces our STGNN approach. In Sec. 4, we discuss the case study conducted on a Swiss PSH plant, including the experimental and training setups. In Sec. 5, we present our results. Finally, Sec. 6 concludes this work and outlines future steps.

2. Background and Related Work

Conventional machine learning applied to power systems have primarily focussed on linear regression models and recurrent neural networks (Zheng, Xu, Zhang, & Li, 2017). These methodologies continue to be effective and provide competitive results, particularly in areas like short-term load forecasting (Guo, Che, Shahidehpour, & Wan, 2021) and daily peak-energy demand forecasting (Kim, Jeong, & Kim, 2022). Since their introduction, GNNs (Bronstein, Bruna, LeCun, Szlam, & Vandergheynst, 2017) have been applied to many tasks in power systems. By now, graph-based deep learning has become a well-established method for analyzing power system data, thanks to its ability to include structural and homophily biases that cannot be modeled conventionally. Message-passing GNNs have been successfully applied to state estimation (Kundacina et al., 2022), and power flow estimations (Ringsquandl et al., 2021). The same tasks have also been addressed using *graph convolutional neural networks*

(GCN) (Fatah, Claessens, & Schoukens, 2021). In the broader power-grid environment, GNNs are also used for wind speed forecasting in renewable energy (Liao, Yang, Wang, & Ren, 2021). Additionally, GNN-based state forecasting was used in a range of downstream tasks in several previous research studies, including graph-based early fault detection for IIoT systems (Zhao & Fink, 2024), anomaly detection in the electrical grid (Li et al., 2022), fault diagnosis for three-phase flow facility (Chen, Liu, Hu, & Ding, 2021), predicting dynamical grid stability (Nauck et al., 2022), and physics-informed parameter and state estimations (Pagnier & Chertkov, 2021b, 2021a).

Other works have used spatial-temporal extensions of GNNs in the electrical domain for residential load-forecasting (Lin et al., 2021), fault diagnostics in power distribution systems (Nguyen, Vu, Nguyen, Panwar, & Hovsapian, 2022), and with complex-value extensions (T. Wu, Scaglione, & Arnold, 2022) for state forecasting. In another lie of research, (Wang et al., 2022) propose spatial-temporal graph learning for power flow analysis, where the graph is dynamically created from thresholded normalized mutual information. At the component feature level, attention-based graph learning (GAT) has been applied to power flow analysis (Jeddi & Shafieezadeh, 2021) and in a different work for probabilistic power flow to quantify uncertainties of distribution power systems (H. Wu, Wang, Xu, & Jia, 2022). To the best of our knowledge, the approach of employing a self-attention mechanism at the graph level to learn a graph structure that integrates electrical and hydraulic data using *graph neural networks* has not yet been addressed in previous research.

In the context of power systems, data fusion represents a crucial technique for enhancing the accuracy and reliability of forecasting algorithms by integrating diverse data sources. In the electrical domain, the fusion of diverse electrical system information was utilized to estimate the voltage in distribution networks (Y. Zhu, Gu, & Li, 2020), using cross-correlations between individual transformers. Another research study achieved state-of-the-art multi-site photovoltaic (PV) power forecasting (Simeunović, Schubnel, Alet, & Carrillo, 2022), fusing spatially distributed PV data by exploiting the intuition that PV systems provide a dense network of virtual weather stations. Integrating weather station data was also explored for anomaly detection for the industrial internet of things (Y. Wu, Dai, & Tang, 2022). Given its importance, modeling the interaction between the PSH subsystems has previously been explored using higher-order numerical simulation (SIMSEN) (Simond, Allenbach, Nicolet, & Avellan, 2006). However, operating numerical simulators in practice requires precise calibration and, consequently, extensive documentation of the components, which is typically not readily available. This calibration step is indispensable due to the components exhibiting highly non-linear characteristics (Nicolet et al., 2007). Given the significant variations in designs across different PSH assets, and the necessity for expert knowledge (which is often unavailable), applying this simulator-based methodology is often infeasible in real-world applications. For PSH, these limitations shift the focus to data-driven interaction modelling with GNNs, which is computationally affordable and does not necessitate expert knowledge, yet remains unexplored.

3. METHODOLOGY

This section introduces the Spectral-Temporal Graph Neural Network (STGNN) that we propose for the fusion of electric and hydraulic data in PSH state forecasting. In Section 3.1, we define the forecasting problem. From Section 3.3 onwards, we decompose the forecasting problem into learning the underlying latent graph structure from time series data (Sec. 3.4). Subsequently, on the learned graph, we introduce graph-spectral and time-spectral filtering (Sec. 3.6). An overview of the methodology is provided in Figure 1.

Notation: In this work, we use slicing notation denoted by the colon (:) symbol. Given a matrix $A \in \mathbb{R}^{m \times n}$, where m and n represent the number of rows and columns respectively, slicing is expressed as $A[i:j, k:l]$ or $A^{i:j,k:l}$. This notation represents the selection of rows i through $j-1$ and columns k through $l-1$ of matrix A. If i or k is omitted, it implies starting from the first row or column, respectively. Similarly, if j or l is omitted, it implies selection until the last row or column, respectively. We use $\otimes$ to denote element wise multiplication and $\oplus$ for concatenation. The Frobenius norm is denoted as $\| \bullet \|_{\mathrm{F}}$.

3.1. Problem Formulation

The specific objective of this work is to provide accurate state forecasting for the electrical subsystem of the PSH. Our approach utilizes learnable graph-spectral and time-spectral filtering to compute state predictions (the forecast). Let $\bar{\mathbf{X}} \in \mathbb{R}^{T \times H+E}$ represent the smoothed time series data $\bar{\mathbf{X}} = S(\mathbf{X})$ computed from unsmoothed time series $\mathbf{X}$ using the smoothing function S for E sensors in the electric and H sensors in the hydraulic subsystem, respectively, over a time period T. We define our forecasting model as a function $M : \mathbb{R}^{w \times H+E} \rightarrow \mathbb{R}^{h \times E}$ for a specific point in time $w < t < T$, that operates on input windows of the data of length w sliced as $\bar{\mathbf{X}}[t-w : t]$. The goal is to forecast the subset of electrical sensors $\bar{\mathbf{X}}[t : t+h, : E]$ for a horizon of size h. Model M operates on a graph $\mathcal{G}$ that is either inferred from the input data by a parameterized function $\mathcal{G}_\phi(\bar{\mathbf{X}}[t-w : t])$, which is trained alongside M, or may be provided apriori. We introduce two sets of learnable parameters: θ for the filtering M_θ and ϕ for the graph learning $\mathcal{G}_\phi$. Thus, state forecasting and state reconstruction estimates denoted by $\widetilde{\bullet}$ at a selected timepoint t, are computed by the model M as follows:

$$\widetilde{\mathbf{X}}_{\text{elec}}^{t:t+h}, \widetilde{\mathbf{X}}_{\text{elec}}^{t-w:t} = M_\theta(\bar{\mathbf{X}}^{t-w:t} \mid \mathcal{G}_\phi) \tag{1}$$

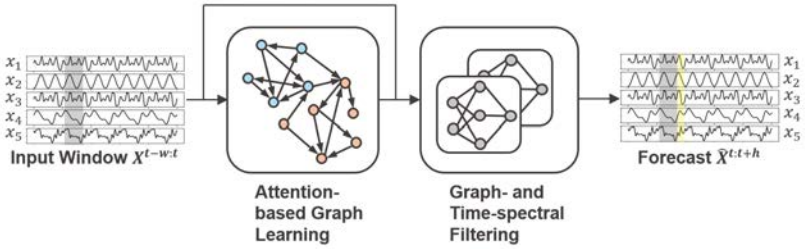


Figure 1. An overview of the two processing steps of our *spectral-temporal graph neural network* to fuse data from the electrical and hydraulic subsystems of a PSH. Utilizing *attention-based graph learning*, our method dynamically constructs a graph based on an input window $\bar{\mathbf{X}}^{t-w:t}$. Subsequently, by operating on this graph, the *graph- and time-spectral filtering* module efficiently extracts information from the hydraulic and electrical sensor data to forecast the subset of electrical sensors $\widetilde{\mathbf{X}}_{\text{elec}}^{t:t+h}$.

3.2. Training Objective

By employing a sliding window approach, we construct a training dataset $\mathcal{X}^{\text{Train}}$ of length N_{train}, where N_{train} depends on the training data split. We also construct analogous validation and test datasets $\mathcal{X}^{\text{Val}}$ and $\mathcal{X}^{\text{Test}}$, respectively:

$$\mathcal{X}^{\text{Train}} = \{\bar{\mathbf{X}}_{\text{elec}}^{t:t+h}, \bar{\mathbf{X}}_{\text{elec}}^{t-w:t}, \bar{\mathbf{X}}^{t-w:t}\}_{t=w}^{N_{\text{train}}}$$

We optimize the parameters of the model by minimizing the forecasting error $\mathbf{E}_{\text{f}}^t = \widetilde{\mathbf{X}}_{\text{elec}}^{t:t+h} - \bar{\mathbf{X}}_{\text{elec}}^{t:t+h}$. To learn meaningful and compact representations, we introduce an optional reconstruction error $\mathbf{E}_{\text{r}}^t = \widetilde{\mathbf{X}}_{\text{elec}}^{t-w:t} - \bar{\mathbf{X}}_{\text{elec}}^{t-w:t}$ for regularization. The final training objective is expressed as:

$$\mathcal{L}(\widetilde{\mathbf{X}}, \bar{\mathbf{X}}) = \sum_{t=w}^{N_{\text{train}}} \left(\lambda_{\text{f}} \|\mathbf{E}_{\text{f}}^t\|_F^2 + \lambda_{\text{r}} \|\mathbf{E}_{\text{r}}^t\|_F^2\right)$$

During the model's training process, we identify the optimal parameters:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathcal{L}(\widetilde{\mathbf{X}}, \bar{\mathbf{X}}; \phi, \theta) \tag{2}$$

3.3. Node Features and Graph Representation

For the forecasting problem, we consider spatially distributed sensor sites modeled as nodes (vertices) $v \in V$ of a graph that spans the pumped storage hydropower plant. Due to differences in raw sensor sampling rates, we use resampled time series based on simple moving averages, taking into account the true sensor sampling rate of the j-th sensor S_j. This step smooths the time series:

$$\bar{\mathbf{X}}[i,j] = S(\mathbf{X}[i,j]) = \frac{1}{S_j} \sum_{\tau=1}^{S_j} s_\tau^j \tag{3}$$

We model each measurement site, containing one or more sensors as an individual node. Each node is associated with a feature vector $\mathbf{x}_v^{t-w:t} \in \mathbb{R}^{w \times d}, \quad \forall v \in V$, containing a window w of the smoothed sensor data and additional $d-1$ covariate dimensions such as a temporal encoding. This strategy is uniformly applied to both the electrical and hydraulic components within the pumped-storage power plant environment. Nodes are exclusively assigned to one of the sets: $\mathbb{1}_{\text{el}}(v) = 1$ for electrical components or $\mathbb{1}_{\text{hyd}}(v) = 1$ for hydraulic components, ensuring $\mathbb{1}_{\text{hyd}}(v) + \mathbb{1}_{\text{el}}(v) = 1$. As input for the subsequent model M, we consider the joint graph:

$$\mathcal{G}_\phi = \left(V_{\text{el}} \cup V_{\text{hyd}}, E_\phi(\bar{\mathbf{X}}[t-w:t])\right)$$

where the edges may be learned by a parameterized function E_ϕ.

3.4. Attention-based Graph Learning

We define a trainable function that implements self-attention among the sensor nodes to infer the edges $E_\phi(\mathbf{X})$ of the graph $\mathcal{G}$. To compute the self-attention, we first map the time series to an embedding space $\mathbf{E} = \text{GRU}(\bar{\mathbf{X}})$ using a gated recurrent unit (GRU). We then proceed by computing the self-attention of the embedded time series. For this purpose, we define a query sequence Q and a key sequence K to compute the attention scores W:

$$Q = EW^Q, K = EW^K, W = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{4}$$

by linear projection with the trainable matrices $\phi = (W^Q, W^K)$. Unlike in *graph attention networks* (GAT) (H. Wu et al., 2022), which define attention over features of a pre-existing graph, we directly compute the symmetrically normalized graph Laplacian L from the attention scores W , which we convert into a symmetrical adjacency matrix $A = \frac{1}{2}(W + W^T)$. We compute the Laplacian as $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where L is the Laplacian matrix, I is the identity matrix, D is the diagonal degree matrix of A, the adjacency matrix. L is then used for the graph spectral filtering in Section 3.6.

3.5. Spectral-temporal Graph Neural Network

To predict the sensor dynamics, the model processes the input data $\bar{\mathbf{X}}^{t-w:t}$ on the learned graph $\mathcal{G}$ (obtained as introduced in Section 3.4) by mapping the input data from the spatial-temporal vertex domain of the sensor signals to a spectral latent representation. This mapping is achieved through the sequential application of graph-spectral and time-spectral transformations, as introduced in Sec. 3.6. The corresponding inverse transformations are utilized to reconstruct the sensor signal in the spatial-temporal domain. To address the problem of vanishing gradients and performance degradation with increasing network depth, we introduce skip connections to compute the final forecast. We denote the output of the residual blocks as s_k. Thus, model M can be expressed with spectral filtering (F), and a bypass layer f_b as follows in the recursive equation:

$$(s_k^f, s_k^b) = \begin{cases} F\left(\bar{\mathbf{X}}\right) \mid \mathcal{G}_\phi(\bar{\mathbf{X}})\big), s_0^b = \bar{\mathbf{X}} & k = 0 \\ F\left(\sigma(s_{k-1}^b - f_s(s_{k-1}^b)) \mid \mathcal{G}_\phi(\bar{\mathbf{X}})\right), & k > 0 \end{cases} \tag{5}$$

where the final forecast is computed as $\widetilde{\mathbf{X}}_{\text{elec}}^{t:t+h} = \Omega_f\left(\sum_{i=1}^k s_i^f\right)$ with an application-specific feedforward head function (Ω), and the backcast as $\widetilde{\mathbf{X}}_{\text{elec}}^{t-w:t} = \sum_{i=1}^k s_i^b$.

3.6. Graph- and Time-spectral Filtering

For the spectral filtering module F, we use graph convolutional filtering and the trainable *Spe-Seq Cell* S_θ introduced in (Cao, Wang, et al., 2021). We denote the graph Fourier transform as $\mathcal{GF}$, and its inverse as $\mathcal{IGF}$. The j-th channel y_j in the graph-spectral domain is therefore computed as follows:

$$y_j = \mathcal{IGF}\left(\sum_i g_{\theta_{ij}}(\Lambda_i) S_\theta(\mathcal{GF}(\mathbf{X}_i))\right). \tag{6}$$

In the graph-spectral domain, we implement the parameterized filtering $g_\theta(\Lambda_i)$ on the eigenvalues Λ. Instead of computing y_j directly, we compute the Chebyshev polynomials of L to efficiently approximate the graph Fourier transform without performing the costly eigenvalue decomposition of the Laplacian matrix $L = U\Lambda U^T$, where U is the graph's eigenvector matrix. We obtain the i-th Chebyshev polynomial $T_i(\bullet)$ with the recurrence relation: $T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$. Thus, we implement graph-spectral filtering with the graph spectral operator $g(L)$ as follows:

$$g(\widetilde{L})\mathbf{X}_j \approx \sum_{n=0}^{N} c_n S_\theta(T_n(\widetilde{L})\mathbf{X}_j) \tag{7}$$

where c_n are the learnable parameters and $\widetilde{L} = 2L/\lambda_{\max} - I_N$ is the normalized Laplacian matrix. The *Spe-Seq Cell* enhances the output of $\mathcal{GF}$, treating it as a multivariate time-series in the graph-spectral domain. It then elevates this output into the time-spectral domain to learn feature representations. To achieve this, the *Spe-Seq Cell* uses the Discrete Fourier Transform (DFT) and gated linear units (GLU) for element-wise modulation of the signal in time-spectral domain as follows:

$$\text{GLU}(\mathbf{x}) = \mathbf{x} \otimes \sigma(\mathbf{W}^g\mathbf{x} + \mathbf{b}^g) \tag{8}$$

This approach effectively implements convolution on the multivariate time-series in the graph-spectral domain.

4. Case Study and Experimental Setup

The dataset in this case study was obtained in collaboration with the Swiss Federal Railways (SBB). SBB maintains a separate railway traction current network (RTN) that operates at a frequency of 16.66 Hz to power rolling stock across Switzerland. The power plant operators of SBB use *supervisory control and data acquisition (SCADA)* protocols to transmit sensor data to a centralized *energy management system* (EMS). This setup allows for real-time monitoring of assets, ensuring timeliness and synchronicity between sensor signals, making it a technically sound environment to evaluate the proposed methodology.

Objective: For this case study, our aim is to forecast the currents measured by the electrical sensor network of the PSH. From an operator's perspective, forecasting currents is particularly compelling when dealing with rolling stock, given their highly dynamic current profile that is vastly different to residential power grids. While the residential sub-grid of Zurich, the largest City of Switzerland, is subject to transient load changes within 15 minutes intervals of up to 35MW, the RTN of SBB experiences load changes up to 250 MW whithin the same time interval due to the orchestrated and periodic timetable of the Swiss railway network (Halder, 2018). The importance of accurate current forecasts is additionally heightened because electrical components in power systems, like transformers and conductors, have thermal limitations that depend on the amount of current flowing through them. Unlike voltage levels, current levels in the PSH dynamically react to transient loads changes. Anomalies such as sudden increases or decreases in power demand, or failures in equipment, are therefore more immediately reflected in current fluctuations. We therefore focus on phasor current forecasting in this study.

Dataset & Data Preparation: We collected data spanning four months from a PSH in Switzerland, consisting of readings from 58 sensors that monitor pressures, flow rates, and lake levels of the hydraulic subsystem, as well as electrical currents from seven generating units, including connected substations in the electrical subsystem. The time series are averaged to a 1-minute resolution and were collected from

January to March in 2021. We maintained the temporal ordering of training (70% of the data), validation (15%) and test (15%) datasets to ensure that the validation and test indices are sequentially higher than the training indices. We normalize the data using feature-wise min-max scaling. To provide a comprehensive understanding of the dataset, we show a detailed segment of the sensor data in Figure 2.

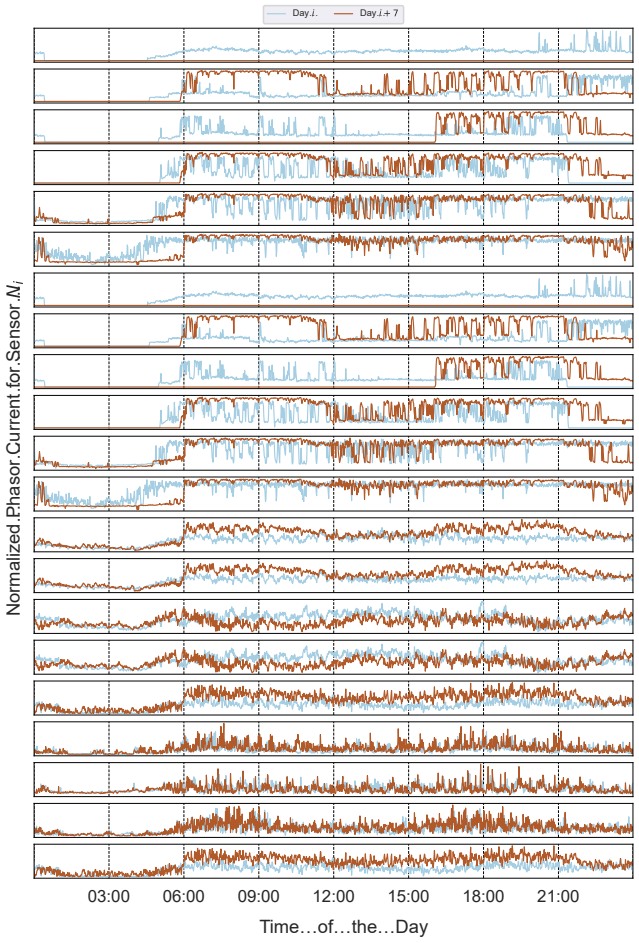


Figure 2. Segment of the dataset, displaying all 21 normalized phasor current sensors (the forecasting target of our case study), indicating the dynamic nature of the sensor measurements. We show the same day of the week (i and $i + 7$) for two consecutive weeks.

Model & Training: The experiments were conducted on an NVIDIA RTX3060 using PyTorch 2.0 and CUDA 11.8 for the development and training of the models. The proposed model utilizes a window size (w) of 24 and a horizon size of 1, meaning that it predicts the currents for the next minute. This configuration is tailored to the synchronized operation of the Swiss railway network, which organizes its periodic timetable in half-hourly intervals. Our selected model's input window takes this operational profile into account, thereby reducing the influence of the previous interval. During model fine-tuning, we truncate the Chebychev polynomial expansion to $k = 4$ for both the graph- and time-spectral filtering. We set the number of residual blocks to two and configure the *Spe-Seq Cells* to five layers. We adjust the negative slope of *LeakyReLU* α to 0.2, and dropout to 0.5 for regularization. We use Adam for optimization.

Table 1. Overview of trainable parameters in the neural network models.

Model	Parameter
MLP (4-layer, el+hy)	2.8 M
STF (el+hy)	366 K
Ours (el)	481k
Ours (el+hy)	481k

4.1. Baseline Approaches for Comparison

We compare our approach with relevant baselines for time series forecasting in the power grid domain, such as a linear model with trend decomposition, a fully connected neural network (FNN) and a *recurrent neural network*, specifically the LSTM. Given the recent increase in attention towards transformer-based energy forecasting, we also consider the *time series transformer* model (*Spacetimeformer*, (Grigsby, Wang, & Qi, 2022), denoted as STF) Additionally, we include an Attention-based GNN (A3-GCN, (J. Zhu, Song, Zhao, & Li, 2020)) that has been demonstrated to outperform the classical GCN on similar forecasting tasks. All baseline models were trained and validated on the same dataset and with the same input window size. Each model was trained to reach convergence on the validation dataset.

5. Results

In this section, we summarize the numerical results to evaluate the proposed method and compare it to the baselines. Additionally, we study the output from the graph learning and compare it to the connectivity of the PSH asset. Furthermore, we ablate the hydraulic information from the model input to assess its benefits. All results are based on the dataset introduced in Section 4.

In the initial step, we compare the performance of the adopted *spectral-temporal graph neural network* to the baseline models introduced in Section 4.1 based on normalized mean squared error (NMSE). We summarize model performances in Table 2. Our model surpasses all conventional baselines, including LSTM (by 28%) and the, in terms of parameters, much larger FNN (by 14%), as evaluated by the NMSE. A3-GCN is unable to integrate hydraulic information due to the lack of a graph. Spacetimeformer (STF) can learn from hydraulic signals but does not outperform our method.

For the evaluation of A3-GCN, we translate the PSH's electrical network diagram into a processable graph, as having apri-

Table 2. Average (normalized) model performance across nodes, comparing six different methods. We indicate whether the PSH network diagrams were translated into a processable graph for the computation (Network Diagram) and epmhasize if hydraulic (Hyd.) or electric (El.) information was used for training.

Method	El.	Hyd.	Network Diagram	Type	NMSE
Linear	✓	✓	✗	-	1.11e-1
A3-GCN	✓	✗	✓	GCN	8.74e-3
LSTM	✓	✓	✗	RNN	7.51e-3
MLP (3-layer)	✓	✓	✗	FNN	6.84e-3
MLP (4-layer)	✓	✓	✗	FNN	6.21e-3
STF	✓	✗	✗	Transformer	5.84e-3
STF	✓	✓	✗	Transformer	5.83e-3
Ours	✓	✗	✗	Att. GCN	<u>5.71e-3</u>
Ours	✓	✓	✗	Att. GCN	**5.34e-3**

ori graph is a computational requirement for the method. Surprisingly, we found that the A3-GCN is outperformed by the much simpler LSTM by 14% in terms of NMSE. This finding highlights that the intuitive approach of applying GNN directly to a graph derived from schematic diagrams, does not always yield acceptable results. In this context, since our proposed STGNN is also based on GCN, the 34.7% improvement in NMSE illustrates that, beyond choosing the right model, finding a suitable computational graph is crucial for processing PSH data. Our results provide further support for the observation in (Ringsquandl et al., 2021) that statistical properties of graphs derived from network diagrams of power grids may be unsuitable for direct graph processing. Graphs derived from such network diagrams significantly differ from those typically discussed in the graph-theoretical literature, with statistical properties like lower clustering-coefficients, lower node degrees, and higher graph diameters, which could explain the subpar performance of A3-GCN in the state forecasting task. From a message-passing perspective, the specific properties of these graphs hinder effective message propagation unless the GNN comprises many layers. Unfortunaltely, this model choice, in turn, significantly boosts oversmoothing, which is already a prevalent challenge in the power grid environment due to the high similarity of the electrical sensor data.

We assess the performance of time-series transformers (Spacetimeformer), which are structurally similar to our attention-based GNN approach, because they incorporate a self-attention layer across the one-dimensional timeseries. However, the experiment with Spacetimeformer displays a 9.2% reduction in performance in terms of NMSE compared to the STGNN. Additionally, we found it difficult to scale Spacetimeformer to the problem without overfitting. Altrough showcasting respectable performance, compared to the STGNN, STF did not benefit from the additional information from hydraulic systems (resulting in a 0.1% improvement).

An advantage of our STGNN, compared to conventional methodology suited for multivariate time series analysis such as LSTM, is that we have access to the learned graph topology. To derive insights from the inferred graph, we calculate the mean attention $a_{ij} = \frac{1}{N}\sum_{i=1}^{N} a_{ij}$ over the test data set. We expect our method to converge to the same graph for randomized training initialization when learning physical relationships between the sensors. To verify our expectation, we visually compare the average attention across random seeds in Figure 4. Additionally, our analysis reveals that the inferred attention graph's minimal parameterization yields temporally stable graphs, accurately reflecting the situation in the PSH, which usually has stable topology across time.

Interestingly, the attention graph recovers casual relations given by the functioning of the hydropower plant and therefore shows similarity to the underlying physical network of the hydropower plant. In Figure 6, we depict the relationships between the water inflow (flow-rates, pressures), the generator units (groups and transformers, indicated by TRF), and the PSH outlets (substations, denoted by SP), overall making the model's predictions more interpretable. Leveraging this interpretability, we compare Figures 5 and 6. Surprisingly, our model focuses on the PSH outlets to predict the power plant input's phasor currents in the absence of hydraulic information from the forecast, which could explain the more severe outliers in the forecast (Figure 3). When adding hydraulic information to the forecast, we observe that the model makes additional use of penstock flow-rate and pressure sensor data, thereby improving the prediction quality.

5.1. Ablation Study

In an ablation study, we exclude the hydraulic sensor data from the forecast to validate the effectiveness of fusing electric and hydraulic sensor data for improving the electrical state forecasting. We find that incorporating the hydraulic subsystem leads to an 6.5% reduction in NMSE. The NMSE absolute forecasting performance from the ablation experiment is included in Table 2. Figure 7 and 8 evaluate the relative improvement on a per-node basis, demonstrating that the benefits of our methodology are distributed across most sen-

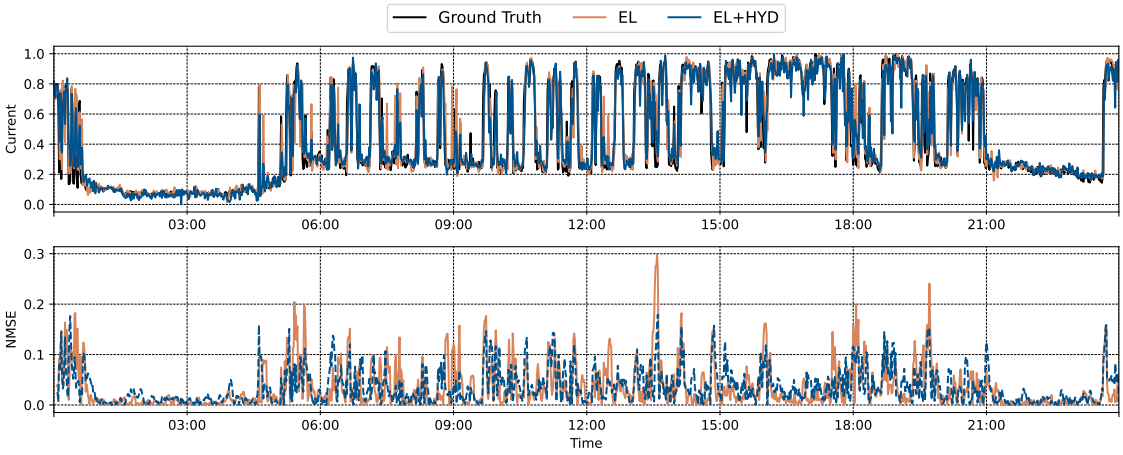


Figure 3. Comparison of normalized phasor current forecasts with (EL+HYD) and without the hydraulic information (EL) for our proposed STGNN model. We show the forecast for single node N_i ($i = 1$) across a randomly selected day including ground truth. In the upper Figure, we display the dynamic range of the forecast. In the lower Figure, we display the normalized MSE of both approaches with respect to the ground truth. Removing hydraulic information results in heightened discrepancies and more pronounced outliers in the predictions. First, we select the data based on the above criteria. Then, we normalize the selected data using min-max scaling.

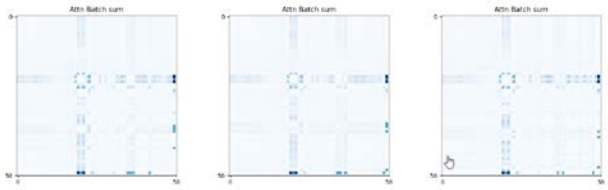


Figure 4. The averaged learned attention across the test set of the *attention-based graph learning* module over all 58 signals from the electrical and hydraulic subsystems. We show three random seeds. The learned attention is stable for different random initializations.

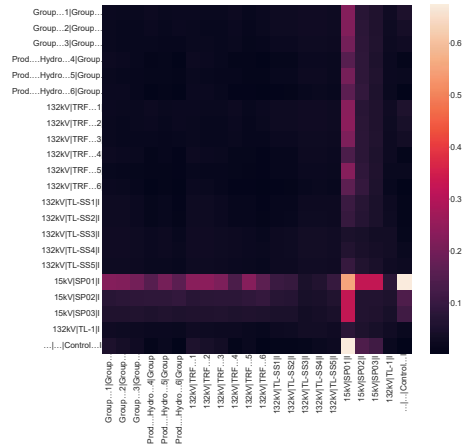


Figure 5. The heatmap represents the averaged learned attention by the *attention-based graph learning* module across the test set as for the model processing only electrial information (EL). Counterintuitively, the model focuses on the PSH output (SP) when forecasting the phasor currents of the electromagnetic generators, which represent the PSH input.

sor forecasts (nodes), thereby ensuring that no sensor forecast experiences a major decline in performance from including hydraulic sensor information.

6. Conclusions

In this paper, we demonstrate that integrating information across the electrical and hydraulic subsystems is beneficial for state forecasting in pumped-storage hydropower plants (PSH). Our proposed *spectral-temporal graph neural network* is the first approach to integrate information across the PSH's subsystems by applying *attention-based graph learning*, which effectively represents PSH states for short-term phasor current forecasting. Compared to numerical simulation, our method requires neither knowledge of the underlying network and sensor connectivity graph nor a tedious calibration step. Through a real world case study, we demonstrate that relying exclusively on graphs derived from network diagrams for state forecasting does not always yield the best performance. We highlight the advantages of learning a PSH-wide graph, complementing the critical perspective on network-diagram-derived graphs introduced in (Ringsquandl et al., 2021). Moreover, we show that our method remains interpretable, unlike other deep-learning methods that process electrical and hydraulic data simultaneously. Future work looks to reintegrate the underlying network diagram while maintaining the flexibility of attention-based graph learning, thereby harnessing the strengths of both approaches. This could also allow for the incorporation of physics-informed losses, such as electromagnetic generator efficiency or power flow, which may reduce the volume of training data required.

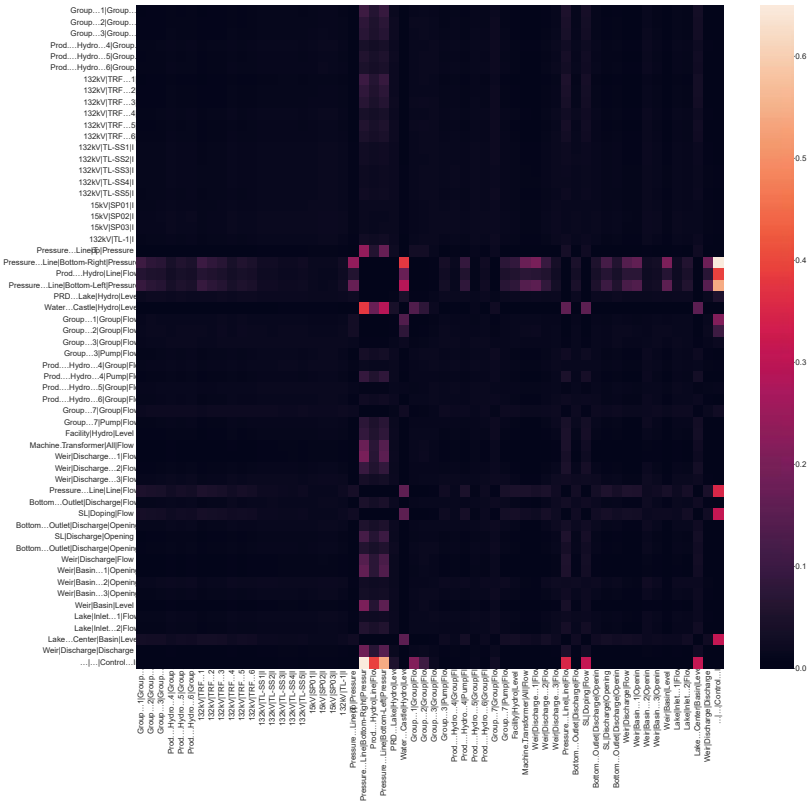


Figure 6. The heatmap visualizes the averaged learned attention of the *attention-based graph learning* module across the test set visualized for the model that processes both electrial and hydraulic information (EL+HYD). Notably, the model focuses on the hydraulic subsystem (the PSH input) when forecasting the phasor currents of the electromagnetic generators.

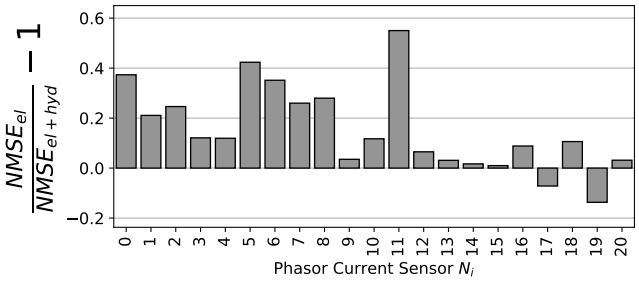


Figure 7. Relative improvements across the test set (normalized MSE, averaged) for our STGNN models with (EL+HYD) and without hydraulic information (EL). for each of the 21 phasor currents sensors of the electric subsystem. 19 out of 21 phasor current forecasts are improved by the additional hydraulic information.

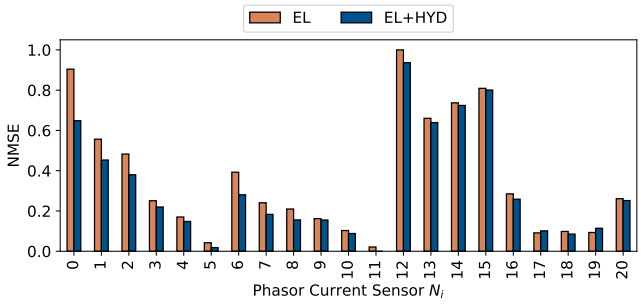


Figure 8. The normalized MSE averaged across the test set is displayed, comparing our STGNN models with (EL+HYD) and without hydraulic information (EL). The performance across the 21 phasor currents sensors of the electric subsystem is shown. The additional hydraulic information improves the phasor current forecasts in 19 out of the 21 cases.

REFERENCES

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017, July). Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42. doi: 10.1109/MSP.2017.2693418

Cao, D., Li, J., Ma, H., & Tomizuka, M. (2021, May). Spectral Temporal Graph Neural Network for Trajectory Prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1839–1845). doi: 10.1109/ICRA48506.2021.9561461

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., … Zhang, Q. (2021, March). *Spectral Temporal Graph Neural Network for Multivariate Time-*

series Forecasting (No. arXiv:2103.07719). arXiv. doi: 10.48550/arXiv.2103.07719

Chen, D., Liu, R., Hu, Q., & Ding, S. X. (2021). Interaction-Aware Graph Neural Networks for Fault Diagnosis of Complex Industrial Processes. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14. doi: 10.1109/TNNLS.2021.3132376

Fatah, M. G. A., Claessens, B. J., & Schoukens, M. (2021, August). Integrating Power Grid Topology in Graph Neural Networks for Power Flow. *Eindhoven University of Technology*, 11.

Grigsby, J., Wang, Z., & Qi, Y. (2022, May). *Long-Range Transformers for Dynamic Spatiotemporal Forecasting* (No. arXiv:2109.12218). arXiv.

Guo, W., Che, L., Shahidehpour, M., & Wan, X. (2021, January). Machine-Learning based methods in short-term load forecasting. *The Electricity Journal*, *34*(1), 106884. doi: 10.1016/j.tej.2020.106884

Halder, M. (2018, September). *Power Demand Management – Smart Grid @ SBB*. Innotrans.

Jeddi, A. B., & Shafieezadeh, A. (2021, December). A Physics-Informed Graph Attention-based Approach for Power Flow Analysis. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1634–1640). doi: 10.1109/ICMLA52953.2021.00261

Kim, H., Jeong, J., & Kim, C. (2022, November). Daily Peak-Electricity-Demand Forecasting Based on Residual Long Short-Term Network. *Mathematics*, *10*(23), 4486. doi: 10.3390/math10234486

Kundacina, O., Cosovic, M., & Vukobratovic, D. (2022, April). *State Estimation in Electric Power Systems Leveraging Graph Neural Networks* (No. arXiv:2201.04056). arXiv.

Li, S., Pandey, A., Hooi, B., Faloutsos, C., & Pileggi, L. (2022, September). Dynamic Graph-Based Anomaly Detection in the Electrical Grid. *IEEE Transactions on Power Systems*, *37*(5), 3408–3422. doi: 10.1109/TPWRS.2021.3132852

Liao, W., Bak-Jensen, B., Radhakrishna Pillai, J., Wang, Y., & Wang, Y. (2022). A Review of Graph Neural Networks and Their Applications in Power Systems. *Journal of Modern Power Systems and Clean Energy*, *10*(2), 345–360. doi: 10.35833/MPCE.2021.000058

Liao, W., Yang, D., Wang, Y., & Ren, X. (2021, March). Fault diagnosis of power transformers using graph convolutional network. *CSEE Journal of Power and Energy Systems*, *7*(2), 241–249. doi: 10.17775/CSEEJPES.2020.04120

Lin, W., Wu, D., & Boulet, B. (2021, November). Spatial-Temporal Residential Short-Term Load Forecasting via Graph Neural Networks. *IEEE Transactions on Smart Grid*, *12*(6), 5373–5384. doi: 10.1109/TSG.2021.3093515

Ma, Y., Liu, X., Shah, N., & Tang, J. (2023, July). *Is Homophily a Necessity for Graph Neural Networks?* (No. arXiv:2106.06134). arXiv.

Nauck, C., Lindner, M., Schürholt, K., Zhang, H., Schultz, P., Kurths, J., … Hellmann, F. (2022, April). Predicting Basin Stability of Power Grids using Graph Neural Networks. *New Journal of Physics*, *24*(4), 043041. doi: 10.1088/1367-2630/ac54c9

Nguyen, B., Vu, T., Nguyen, T.-T., Panwar, M., & Hovsapian, R. (2022, October). *Spatial-Temporal Recurrent Graph Neural Networks for Fault Diagnostics in Power Distribution Systems* (No. arXiv:2210.15177). arXiv. doi: 10.48550/arXiv.2210.15177

Nicolet, C., Greiveldinger, B., Herou, J. J., Kawkabani, B., Allenbach, P., Simond, J.-J., & Avellan, F. (2007, November). High-Order Modeling of Hydraulic Power Plant in Islanded Power Network. *IEEE Transactions on Power Systems*, *22*(4), 1870–1880. doi: 10.1109/TPWRS.2007.907348

Pagnier, L., & Chertkov, M. (2021a, March). *Embedding Power Flow into Machine Learning for Parameter and State Estimation* (No. arXiv:2103.14251). arXiv. doi: 10.48550/arXiv.2103.14251

Pagnier, L., & Chertkov, M. (2021b, February). *Physics-Informed Graphical Neural Network for Parameter & State Estimations in Power Systems* (No. arXiv:2102.06349). arXiv. doi: 10.48550/arXiv.2102.06349

Ringsquandl, M., Sellami, H., Hildebrandt, M., Beyer, D., Henselmeyer, S., Weber, S., & Joblin, M. (2021, October). Power to the Relational Inductive Bias: Graph Neural Networks in Electrical Power Grids. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 1538–1547). doi: 10.1145/3459637.3482464

Simeunović, J., Schubnel, B., Alet, P.-J., & Carrillo, R. E. (2022, April). Spatio-Temporal Graph Neural Networks for Multi-Site PV Power Forecasting. *IEEE Transactions on Sustainable Energy*, *13*(2), 1210–1220. doi: 10.1109/TSTE.2021.3125200

Simond, J.-J., Allenbach, P., Nicolet, C., & Avellan, F. (Eds.). (2006). Simulation tool linking hydroelectric production sites and electrical networks. *Proceedings of 27th Int. Conf. on Electrical Machines, ICEM*.

Wang, F., Chen, P., Zhen, Z., Yin, R., Cao, C., Zhang, Y., & Duić, N. (2022, October). Dynamic spatio-temporal correlation and hierarchical directed graph structure based ultra-short-term wind farm cluster power forecasting method. *Applied Energy*, *323*, 119579. doi: 10.1016/j.apenergy.2022.119579

Wu, H., Wang, M., Xu, Z., & Jia, Y. (2022, November). Graph Attention Enabled Convolutional Network for Distribution System Probabilistic Power Flow. *IEEE Transactions on Industry Applications*, *58*(6), 7068–

7078. doi: 10.1109/TIA.2022.3202159

Wu, T., Scaglione, A., & Arnold, D. (2022, September). *Complex-Value Spatio-temporal Graph Convolutional Neural Networks and its Applications to Electric Power Systems AI* (No. arXiv:2208.08485). arXiv. doi: 10.48550/arXiv.2208.08485

Wu, Y., Dai, H.-N., & Tang, H. (2022, June). Graph Neural Networks for Anomaly Detection in Industrial Internet of Things. *IEEE Internet of Things Journal*, *9*(12), 9214–9231. doi: 10.1109/JIOT.2021.3094295

Zhao, M., & Fink, O. (2024, January). *DyEdgeGAT: Dynamic Edge via Graph Attention for Early Fault Detection in IIoT Systems* (No. arXiv:2307.03761). arXiv. doi: 10.48550/arXiv.2307.03761

Zheng, J., Xu, C., Zhang, Z., & Li, X. (2017, March). Electric load forecasting in smart grids using Long-Short-Term-Memory based Recurrent Neural Network. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)* (pp. 1–6). doi: 10.1109/CISS.2017.7926112

Zhu, J., Song, Y., Zhao, L., & Li, H. (2020, June). *A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting* (No. arXiv:2006.11583). arXiv. doi: 10.48550/arXiv.2006.11583

Zhu, Y., Gu, C., & Li, F. (2020, July). Cross-Domain Data Fusion On Distribution Network Voltage Estimation with D-S Evidence Theory. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6). doi: 10.1109/IJCNN48605.2020.9207414

Health-aware Control for Health Management of Lithium-ion Battery in a V2G Scenario

Mônica S. Félix[1], John J. Martinez-Molina[2], Christophe Bérenguer[3], Chetan S. Kulkarni[4] and Marcos E. Orchard[5]

[1,2,3] *Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France*
monica.spinola-felix@grenoble-inp.fr
john.martinez@grenoble-inp.fr
christophe.berenguer@grenoble-inp.fr

[4] *NASA Ames Research Center (KBR, Inc), Moffett Field, CA 94043, USA*
chetan.s.kulkarni@nasa.gov
[5] *CASE - Univ. of Chile, Santiago, Chile*
morchard@u.uchile.cl

ABSTRACT

In response to the urgent need to combat climate change and reduce greenhouse gas emissions, the transition towards renewable energy sources such as solar and wind power is indispensable. However, the intermittent nature of these sources poses significant challenges to the stability of power grids. Battery Energy Storage Systems (BESS) offer a viable solution, and there is potential for Electric Vehicles (EVs) to serve as energy reservoirs, thereby bolstering grid stability through Vehicle-to-Grid (V2G) technology. While V2G holds promise, concerns persist regarding the longevity of batteries, particularly with the additional demand from charging and discharging cycles. To address these concerns, this study introduces a health-aware control strategy for V2G service scenarios. By employing feedback control mechanisms to adjust degradation rates, the strategy aims to effectively manage battery aging. Simulation outcomes of a V2G scenario with random input sources illustrate the efficacy of this proposed approach, demonstrating its potential applicability in practical settings where battery health needs to be managed. In summary, this research contributes to the advancement of health-aware strategies for an interconnected grid where electric vehicles participate as energy sources, with a primary focus on optimizing battery health management while fulfilling grid demands. Future efforts will concentrate on refining optimization strategies and integrating control methodologies with state estimators to ensure the performance of the approach on embedded battery health management systems.

1. INTRODUCTION

In the face of climate change and the urgent need to reduce global greenhouse gas (GHG) emissions, the transition to non-fossil fuels and renewable energy sources is crucial. While photovoltaics and wind power energy are promising solutions, their intermittent nature poses a challenge to the stability of the power grid. In solving this problem, Battery Energy Storage Systems (BESS) are proving to be a crucial component in ensuring a consistent energy supply. In parallel, the proliferation of Electric Vehicles (EVs) offers the opportunity to use their batteries as energy storage units, which can act as an energy buffer during the day reinforcing the stability of the power grid.

The concept of using EV as energy storage known as Vehicle-to-Grid (V2G) offers advantages, but also pose some challenges. According to (Didier et al., 2021) a fleet of 15% electrified cars in France in 2030 would mean an energy stock of 25GWh throughout the day, equivalent to 20% of the daily average production of the French renewable energy grid in 2020. Another notable benefit is the potential for users to recoup their investment by participating in grid-level demand response programs. However, the use of batteries, particularly EV batteries, raises concerns about their longevity as these batteries are more often used in a ”two-shifts” operation. The impact on battery life has direct financial and environmental implications and therefore justify efforts to find V2G strategies that take battery health into account.

A well- known solution is to use the grid operator as an intelligent conductor, requesting energy from multiple energy sources to ensure the effectiveness of grid distribution while reducing the cost of multiple sources, including the cost of battery aging. An overview of such an approach that uses optimized scheduling methods to control the power grid including V2G application is discussed in (Collath, Tepe, Englberger, Jossen, & Hesse, 2022). However, it is important to

note some limitations of this solution, such as a generalized modeling of battery degradation and the non-consideration of aging-related changes in electrical behavior. An additional challenge lies in the prediction of degradation behavior, a task that remains difficult even with the improved accuracy of data-driven methods.

With recent advancements in online estimation methods for determining the state of health in individual battery packs, the feasibility of health management has expanded to the battery management system (BMS) level. The overall objective is to ensure optimal discharge using adaptive control algorithms to alleviate stress factors and manage the aging process throughout the operational lifespan of electric vehicles, accounting for varying conditions. Recent studies have shown the benefits of implementing a health management controller in wind turbines (Kipchirchir, Do, Njiri, & Söffker, 2023). Using a feedback controller framework can effectively mitigate the degradation process and regulate the end-of-life of these systems using control laws and dynamic models to adapt or reconfigure operational processes. It is expected that incorporating this approach into battery-powered applications can also bring benefits as the aging process in can be efficiently managed.

In this sense, this article presents a health-aware control (HAC) strategy to address battery aging by considering the dynamical interaction between operational and stress variables (e.g. state of charge and temperature). The approach is based on the modulation of the degradation-rate using a feedback controller, as proposed in (Félix, Martinez, & Bérenguer, 2023), in a V2G scenario. For this purpose, a novel dynamic model is first proposed that models the degradation-rate as a function of identified stress factors in response to operational demands. As presented in (Pelletier, Jabali, Laporte, & Veneroni, 2017), the stress factors and effects in battery aging are closely interrelated and an optimal control behavior is not obvious. In addition, the discharge process does not behave linearly and suffers from the fluctuations of aging mechanisms. Therefore, based on the proposed model, a control design is also presented that incorporates robust techniques to handle uncertainties inherent in the degradation modeling and randomness induced by the system operation.

The effectiveness of the approach is evaluated through simulations with a degraded battery model that simulates electrical and thermal dynamics, taking into account variations in critical factors such as increments on the internal resistance and reduction on the battery capacity induced by the ageing process and affecting battery autonomy. To demonstrate practical applicability, the article includes a case study of a simulation of in V2G scenario, integrating uncertainties and random elements to highlight the advantages of the approach in real-life. The results initiate a discussion on the benefits of the approach and its limitations of implementing HAC to V2G and further applications.

Accordingly, Section 2 presents the electrical circuit model of a battery subject to degradation and the proposed model of aging behavior. Section 3 describes the design of a feedback control approach to regulate the degradation-rate. Section 4 shows the results obtained by implementing such a health-aware controller at the BMS level of a V2G application. Conclusions and future perspectives are discussed in Section 5.

2. System model of a degraded battery

2.1. Equivalent circuit model

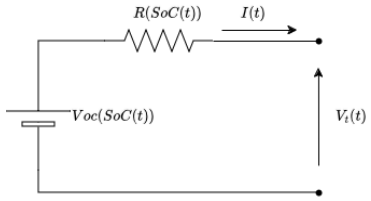


Figure 1. A simplified equivalent circuit model of the battery.

Figure 1 illustrates an equivalent circuit model of a lithium battery simplifies the complex electrochemical processes within the battery into a basic electrical behavior (Pelletier et al., 2017). It includes a voltage source $V_{oc}(k)$ representing the open-circuit voltage, internal resistance R to account for losses within the battery, both dependent of the State-of-Charge (SoC) usually expressed by a parameter $SoC(k)$ to track the available capacity of the battery. This model also incorporates a capacity element C modeling the battery's charge storage capability, and the flow of current $i(t)$ through the battery during charge-discharge cycles. Let us represent such equivalent circuit model as follows:

$$SoC(k+1) = SoC(k) - T_s \frac{I(k)}{(3600 \cdot C_n)} 100 \cdot \gamma(k) \ , \quad (1)$$

$$V_t(k+1) = aV_t(k) + (1-a)E(k) \ , \quad (2)$$

$$\text{with } E(k) = V_{oc}(SoC(k)) - R_n(SoC(k)) \cdot I(k) \cdot \gamma(k) \ . \quad (3)$$

In such model the dynamics of SoC is modeled using Coulomb counting of Eq.1, a function of the charge-discharge current rate $I(k)$ and counting sampling T_s, whereas the behavior of terminal voltage $V_t(k)$ follows Eq. 2 that is driven by a filter parameter a, and V_{oc} and R_n that are functions of the current $SoC(k)$.

2.2. Aging mechanism

In the equivalent circuit model, the state-of-charge and terminal voltage fluctuate in response to the current rate $I(k)$ ($I(k) < 0$ during discharging). Moreover, the charge-discharge behavior is intricately linked to the battery's aging process.

While aging stems from physical-chemical factors, its effects are reflected in electrical characteristics such as capacity fading and increasing internal resistance (Barré et al., 2013). This relationship suggests that these characteristics are influenced by an aging parameter, denoted here as γ. Specifically, the capacity $C(k)$ decrease and resistance $R(k)$ increase can be expressed as functions of γ:

$$C(k) = \frac{C_n}{\gamma(k)} \tag{4}$$

$$R(k) = R_n(SoC(k)) \cdot \gamma(k) \ , \tag{5}$$

where C_n and R_n are the nominal value for capacity and internal resistance.

For simplicity, let us consider γ to be equivalent in both effects, as they pertain to the same aging process. This framework sets the stage for developing a degraded battery model, just as introduced in (Martinez, Félix, Kulkarni, Orchard, & Bérenguer, 2024), where $\gamma = 1$ represents a new battery. During each discharging and charging mission, γ tends to increase until it reaches a maximum value of $\gamma = 2$, indicating that the battery can no longer operate.

2.3. Degradation extended model

As an electrochemical process, charge-discharge behavior incurs energy losses and generates thermal effects. This behavior is externally influenced by ambient temperature $T_{\text{amb}}(k)$ and the Joule effect, which produces heat ($T_{\text{Joule}}(k)$) when the current rate is non-zero. The thermal model can be described by Eq. 6, where c_0 is the inertial parameter of the thermal behavior, and $T_{\text{Joule}}(k)$ is defined by Eq. 7, incorporating factors such as c_1, $R_n(SoC(k))$, $I(k)$, and $\gamma(k)$.

$$T(k+1) = c_0 T(k) + (1 - c_0)\left(T_{Joule}(k) + T_{amb}(k)\right), \tag{6}$$

$$T_{Joule}(k) = c_1 R_n(SoC(k)) I(k)^2 \gamma(k) \tag{7}$$

How γ increases is an important research topic. In this matter, it is known that the increased current generates heat and accelerate degradation by promoting the dissolution of the electrode material and the breakdown of the electrolyte, thus we propose a model for the increase in γ based on the same influences as heating:

$$\gamma(k+1) = c_3 R_n(SoC(k)) I(k)^2 \gamma(k) \tag{8}$$

While this model remains an assumption and approximation, it is crucial for simulating the aging process responsible for increased resistance and decreased capacity. A similar approach to finding a degradation growth model is presented in (Brown et al., 2009) for electro-mechanical actuator applications. Also, this sheds light on which variables of the charging-discharging process could be considered as relevant factors for making decisions regarding aging acceleration.

Note that as the model is posed, an increase in γ increases the acceleration of the aging process itself, similar to how increased temperatures in degraded batteries increase the degradation of the batteries themselves. In addition, a significant increase in γ can lead to instability of T.

3. PROBLEM FORMULATION

This work focus on the application of health-aware discharging for power sale to the grid, known as V2G or V2Market. In this application, there are three main concerns:

1. Supplying the grid with enough energy stored in the battery in order to stabilize it.
2. Monetizing the time the car spends parked in the parking lot without charging.
3. Yet, taking into account the cost of battery degradation since discharging counts as a cycle in the battery's lifespan.

As explained in (Reniers, Mulder, Ober-Blöbaum, & Howey, 2018), when purchasing stored energy, the grid offers a value per kWh correlated with intermittent sources (e.g., solar and wind) availability. Energy prices fluctuate stochastically due to the stochastic behavior of these resources. A grid operator manages participation percentages and demand for each source to optimize production, considering costs, including batteries degradation. This optimization is reviewed in (Collath et al., 2022). While grid demand is generated, users aim to monetize their parked time without charging. Such profitability can be determined by:

$$R = \sum_{k=0}^{t_f} P(k) \times \text{Price}(k) \tag{9}$$

Here, $P(k)$ represents the power sold at discrete time step k, which can be measured by:

$$P(k) = V_t(k) \times I(k) \tag{10}$$

With power demand and price previously established, maximizing profitability becomes a matter of ensuring that $P(k)$ closely matches the power demand and maximizing the discharging interval t_f.

Despite the potential for monetizing a parked car, it is important to acknowledge that there is a cost associated with battery degradation when discharging. This degradation ultimately shortens the battery's lifespan, leading to decreased

performance and necessitating eventual battery replacement. Let us consider this cost as the accumulated degradation over the discharging mission:

$$D = \gamma(t_f) - \gamma(0) \tag{11}$$

To proceed with a solution, let us consider the following:

- The power demand profile (i.e., $P_{grid}(k)$) is pre-defined on an hourly basis, and the power sold never exceeds the power demand ($P(k) \leq P_{grid}(k)$).
- The battery experiences degradation during discharge, as described by Eq. 8. The initial degradation index γ is estimated online using algorithms such as the one proposed in (Didier et al., 2021) or (Martinez et al., 2024).
- The initial State of Charge (SoC) is known and may be lower than the maximum capacity, while a minimum battery SoC is specified and consistently lower than the initial SoC, as represented by:

$$SoC_{\text{min}} \leq SoC(k) < SoC(0) \leq 100\%$$

- The maximum duration for which the vehicle remains parked, selling energy, is predefined as follows:

$$k \leq t_{\text{max}} \sim U(t_1, t_2)$$

- The terminal voltage remains consistently above a safe minimum ($V_t(k) > V_{\text{min}}$).

Figure 2 illustrates the proposed V2G service scenario with battery health management control acting as a discharging auxiliary system. It is assumed that SoH and SoC estimations are available and provided by a BMS.

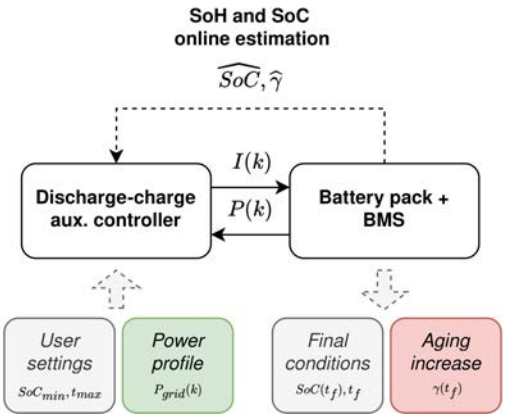


Figure 2. Illustration of Power-sale discharging mission.

3.1. Discharging strategies

The sale of vehicle energy can have numerous strategies. The first and simplest strategy is to manage the delivery of energy to exactly match the grid demand, only stopping the discharge when any of the restrictions (i.e. SoC_{min}, t_{max}, or V_{min}) are triggered. However, this strategy is not optimized, as it does not prioritize maximizing profitability or minimizing degradation costs. Therefore, we can assume other strategies that act on the current rate to maintain power close to demand, while at the same time, avoiding excessive degradation or deliberately degrading to manage battery lifespan. Here we are assuming that the grid is supplied by several other vehicles, and that its stability will not be affected if $P(k) < P_{grid}$.

In the scope of this study, we consider two control strategies:

1. Find the appropriate current $I(k)$ at time k that facilitates discharge in a manner that SoC achieves its minimum by time $tf = t_{max}$, i.e. $SoC(t_{max}) = SoC_{min}$. This strategy optimizes the utilization of parked time and the current, a key factor influencing degradation.
2. Find the optimal current $I(k)$ at time k that allows the degradation rate to follow a predefined reference $\Delta\gamma_{ref}$, ensuring a desired growth rate by time $t_f = t_{max}$, i.e. $\gamma(t_{max}) - \gamma(0) = \Delta\gamma_{ref}$.

The first strategy emphasizes managing the discharge duration. If the discharge time is shorter than the parking duration, the current is adjusted to meet the objective. This approach ultimately influences degradation growth through effective time management.

On the other hand, the second strategy prioritizes degradation effects. It involves tracking a desired γ growth rate either to meet a predefined lifespan or to meet a optimize a value γ^d derived from an optimization problem involving power sold and γ growth rate.

As discussed in (Collath et al., 2022), BESS technology that addresses only issues 1 and 2 of the previously formulated problem does not account for degradation costs over time, reducing the profitability of V2G usage. The two strategies proposed here aim to address the third issue. The first strategy involves mitigating the impact of current on aging increase during discharge, while the second strategy focuses on regulating the degradation-rate by reconfiguring discharge across multiple cycles. Both strategies will inevitably affects the performance of energy delivery to the grid, reducing the profit per discharge. However, this reduction can lead to a significant improvement in battery life. By designing an optimal controller for both strategies, we try to find the best compromise for this trade-off.

3.2. Control framework

For the design of the controller, we consider the system to be controlled written in a discrete-time state-space representation:

$$x_{k+1} = A(\rho_k)x_k + B(\rho_k)u_k + E(\rho_k)d_k\,, \tag{12}$$

with respect to the state vector defined as $x_k \in \Re^n$, the control input variable $u_k \in \Re^m$ and the disturbance input of $d_k \in \Re^p$. The variability of the system is determined by a varying parameter vector ρ_k. When the varying parameter ρ_k is bounded and it belongs to a convex polytopic region Ω_ρ limited by N vertices of the polytopic set $\theta \in \Omega_\rho \subset \Re^L$ defined by ρ_k boundaries. Then, we can write ρ_k as a convex combination of vertices $\theta^{(i)}$ as follows:

$$\rho_k = \sum_{i=1}^{N} \alpha_k^{(i)} \theta^{(i)}, \tag{13}$$

where $\alpha_k^{(i)} \geq 0$ and $\sum_{i=1}^{N} \alpha_k^{(i)} = 1$.

Such a modeling approach is known as polytopic modeling, which enables the construction of robust control designs by guaranteeing stability within the boundaries of the convex set. Since the system is subject to variations due to stochastic disturbances in discharge conditions or changes over time in γ due to aging, this approach will ensure stability for all variable conditions that the discharging process and its dynamics face.

3.2.1. SoC rate control design

According to Eq. 1, SoC decreases dynamically as follows:

$$SoC(k+1) = SoC(k) + SoC'(w(k)) \tag{14}$$

Here, $w(k)$ represents the decision variable that can be adjusted to solve a control problem, specifically a tracking reference problem. To address this problem, we introduce an integrator error tracking $z(k)$ to be minimized. The system to be stabilized is thus defined as:

$$w(k+1) = u(k) \tag{15}$$

$$z(k+1) = z(k) + T_s \cdot (\widehat{SoC}'(w(k)) - SoC'_{ref}(k)) \tag{16}$$

where $u(k)$ represents the control decisions at each sample k, with T_s as the decision rate. $SoC'_{ref}(k)$ is the desired decrease rate of SoC. The linear decrease behavior of $SoC(k)$ imposes a desired rate given by:

$$SoC'_{ref}(k) = \frac{SoC_{min} - \widehat{SoC}(k)}{t_{max} - k} \tag{17}$$

where t_{max} and SoC_{min} are the maximum discharging time and the minimum SoC chosen by the user. $\widehat{SoC}(k)$ can be estimated through online algorithms such as presented in (Didier et al., 2021) with a higher sampling rate.

By choosing the current rate adjustments as the decision variable $w(k)$, we obtain

$$SoC'(w(k)) = \frac{-T_s \cdot 100}{(3600 \cdot C_n)} \gamma(k)(I_{\text{grid}}(k) + w(k)).$$

Now we can define the system matrix as follows:

$$A_k = \begin{bmatrix} 0 & 0 \\ T_s \cdot \rho(k) & 1 \end{bmatrix} \text{ and } B_k = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{18}$$

where $x(k) := [w(k)\ z(k)]$ and $\rho(k) = \frac{-T_s \cdot 100}{(3600 \cdot C_n)} \gamma(k)$, with $\rho(k) \in [\rho_{\text{min}}, \rho_{\text{max}}]$ imposed by minimum and maximum values of γ and nominal capacity.

Finally, we propose to calculate the decisions here using a feedback control law such as:

$$u(k) = -Kx(k). \tag{19}$$

To find the optimal control gain K that minimizes the error z, the system matrices are used to solve a robust Linear-Quadratic Regulator (LQR) problem (see Appendix A) with a Linear Matrix Inequality (LMI) solution.

3.2.2. Aging rate control design

According to Eq. 8, the increase of γ can be expressed as:

$$\gamma(k+1) = \beta(w(k))\gamma(k)$$

In line with the SoC rate control, $w(k)$ denotes the decision variable adjusted to solve a control problem, particularly a tracking reference problem, and an integrator error tracking $z(k)$ is also employed for minimization. The system to be stabilized is thus defined as:

$$w(k+1) = u(k) \tag{20}$$

$$z(k+1) = z(k) + T_s \cdot (\widehat{\beta}(w(k)) - \beta_{ref}(k)), \tag{21}$$

where $\widehat{\beta}(w(k))$ represents the estimated increase rate of γ, T_s is the control decision rate. β_{ref} denotes the current desired increase rate. The exponential growth behavior of γ imposes a desired rate given by:

$$\beta_{ref}(k) = \frac{1}{(t_{max} - k)} \ln\left(\frac{\Delta\gamma_{ref}^{(n)} + \hat{\gamma}(0)}{\hat{\gamma}(k)}\right) \tag{22}$$

where $\hat{\gamma}(0)$ is the estimated γ at the beginning of the cycle and $\hat{\gamma}(k)$ at each control calculation, t_{max} is the imposed maximum discharging interval, and $\Delta\gamma_{ref}$ is the chosen increase increment of the current cycle, which can be calculated in different ways; we propose calculating it based on the value desired of γ for a chosen number of cycles N as follows:

$$\Delta\gamma_{ref}^{(n)} = \frac{1}{2 * (N - (n-1))} \ln\left(\frac{\gamma_{ref}^{max}}{\hat{\gamma}(0)}\right) \tag{23}$$

n represents the current discharge cycle, while γ_{max} denotes the desired level of γ in cycle N. In this scenario, the vehicle is expected to operate in two shifts: during the day in the parking lot and the remainder of the day on regular routes, which counts as an additional full discharge. Although real-

world usage may introduce random degradation during the vehicle's operational shift, we assume that the cumulative effect of these fluctuations does not surpass the degradation equivalent to two full discharges of the vehicle.

According to Eq. 8, the dynamic of β is defined by $\beta(w(k)) = c_3 R_n(SoC(k))(I_{grid}(k) + w(k))^2$. Now, the matrix of the system can be defined as:

$$A_k = \begin{bmatrix} 0 & 0 \\ T_s \cdot \rho(k) & 1 \end{bmatrix} \text{ and } B_k = \begin{bmatrix} 1 \\ 0 \end{bmatrix} , \quad (24)$$

where $x(k) := [w(k)\ z(k)]$ and

$$\rho(k) = 2 * c_3 R_n(SoC(k)) I_{grid}(k),$$

with $\rho(k) \in [\rho_{min}, \rho_{max}]$ imposed by minimum and maximum values of internal resistance and current rate imposed by the grid, and variations on c_3. For this strategy, we also propose utilizing a feedback control law

$$u(k) = -Kx(k),$$

where the control gain K is determined through a robust control LQR problem.

4. Results: V2G scenarios

The degradation battery model presented is used here to simulate an aging battery. The control framework is employed to achieve the objectives of the two different discharge strategies. Firstly, let us describe the simulation scenario used to obtain the results, and then analyze the outcomes of such approaches for battery health management.

4.1. System Description

In real-life scenarios, uncertainties are inherent in the aging process of systems. Various sources of randomness contribute to these uncertainties, stemming from factors such as internal resistance, open circuit voltage, and ambient temperature fluctuations. The interplay of these factors leads to diverse aging acceleration rates, ultimately resulting in varying the battery's lifespan.

The introduced degraded battery model is utilized to simulate the controlled system. The model's parameters, including internal resistance and open-circuit voltage as functions of SoC, are detailed in Appendix B.

The simulated scenario considers the stochastic nature of battery parameters and discharging conditions. The simulation parameters treated as stochastic sources are listed in Table 1.

These conditions vary randomly, as detailed in Table 1, and are subject to change with each simulation of the 20 consecutive days of V2G discharge. For example, Figure 3 outlines the user-defined parameters for each discharge event, including the minimum desired SoC (SoC_{min}) and the maximum discharge duration (t_{max}). These settings collectively determine when the discharge should halt, in conjunction with the minimum voltage, as elaborated in the preceding problem formulation.

Table 1. Simulation parameters

Parameter	Value	Unit
T_{amb}	$\mathcal{N}(23, 3)$	°C
C_1	$\mathcal{N}(2.5, 10^{-2})$	Ah
R_{min}	$\mathcal{N}(0.02, 10^{-2})$	Ω
E_o	$\mathcal{N}(4.2, 10^{-2})$	V
$SoC(0)$	$\mathcal{N}(90, 10)$	%
ϵ_γ	$\mathcal{U}(2e-4, 12e-4)$	-
t_{max}	$\mathcal{N}(4, 1)$	h
SoC_{min}	$\mathcal{N}(10, 1)$	%
P_{grid}	$\mathcal{N}(175, 300)$	W

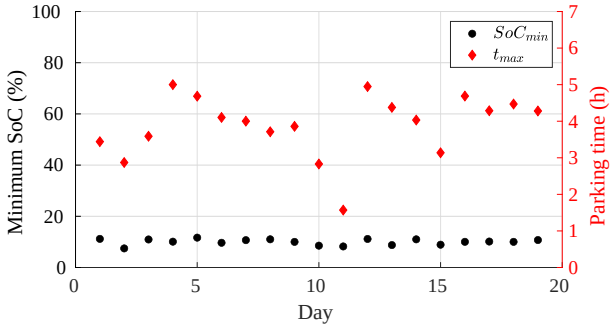


Figure 3. User settings for 20 days in consecutive discharge scenarios.

It is assumed that the battery operates in two shifts: during the day in the parking lot and the rest of the day on regular routes, which contributes to additional aging. Thus, at the start of each new day $n \in [1, N]$, the initial value of $\gamma^n(0)$ is defined as:

$$\gamma^{(n)}(0) = \gamma^{(n-1)}(t_f) + \epsilon_\gamma , \quad (25)$$

where $\gamma^{(n-1)}(t_f)$ represents the level at the end of the last discharging day, and ϵ_γ denotes a random positive additional increase.

Figure 4 displays, then, the initial SoC, $SoC(0)$, representing the battery's starting level upon arrival at the discharge station for the current shift. Additionally, it depicts the additional degradation factor since the last discharge, denoted as ϵ_γ.

Finally, the resulting discharge will depend on the discharging conditions, which will vary with each discharge event. The discharge halts once one of the stop conditions is met. Figure 5 showcases the outcomes of simulations of discharge scenarios without a discharge control. It illustrates diverse discharge histories observed over 20 consecutive days, emphasizing the system's variability influenced by the stochastic nature of battery parameters and discharging conditions.

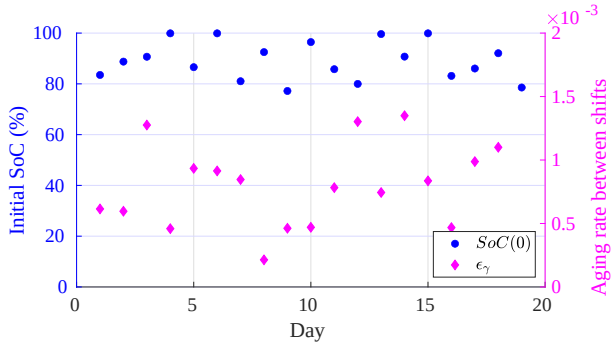


Figure 4. Initial conditions for 20 days in consecutive discharge scenarios.

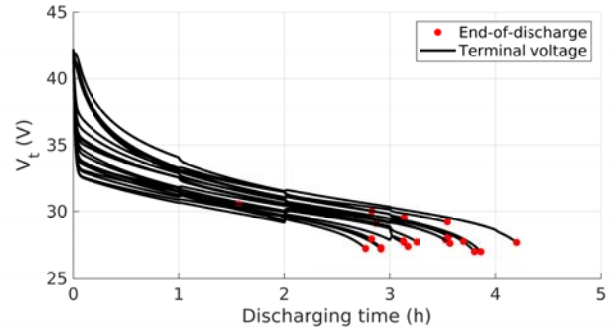


Figure 5. Simulated energy discharge for 20 days in consecutive discharge scenarios for a battery pack of 10 cells.

4.2. Control implementation

In this work, two control strategies are employed to mitigate battery aging, utilizing a feedback control $u(k)$ for the current rate, expressed as $I(k) = I_{\text{grid}} + w(k)$.

SoC Rate Control: For the implementation of the first strategy's control, Eq. (15) is utilized. Employing robust LQR techniques with the previously defined system matrices and simulation parameters, the resulting control gain K is determined as follows:

$$K = [0.011726,\ -70.399]. \tag{26}$$

The initial value for the integral action is calculated according to Eq. (15) as:

$$z_0 = \frac{w_0 + K(1)w_0}{-K(2)}. \tag{27}$$

where the initial decision parameter is chosen to be $w_0 = 0$. Note that, in practice, $w(k)$ is equivalent $u(k-1)$.

Aging rate control: For aging control, Eq. (20) is employed to minimize the reference tracking error. The rate adjustment is accomplished through a feedback control $u = -Kx(k)$, where K is also computed by solving a robust LQR problem with the provided model parameters, yielding the following values:

$$K = [0.9987,\ 5444.5]. \tag{28}$$

It is expected an aging rate reference to be determined according to Eq. (23) using the following parameters:

$$n \in [1, N],\ N = 20,\ \gamma_{\text{ref}}^{\max} = 1.025.$$

Here, $\gamma_{\text{ref}}^{\max}$ is the desired aging parameter value at the conclusion of 20 days, which is set to be lower than the expected value of standard discharging, but could be chosen to respect a prognostic and health management constraint.

4.3. Simulation results

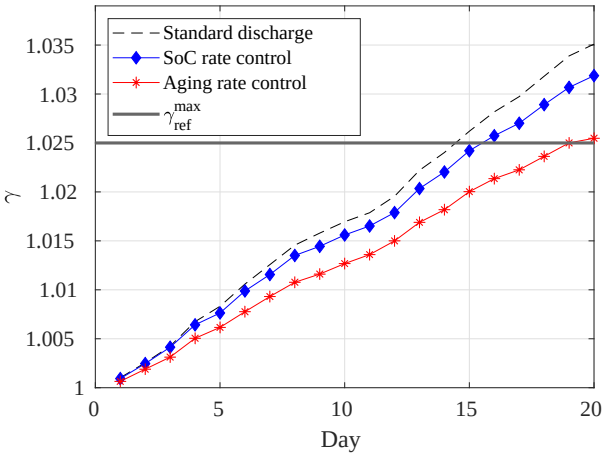
When implementing the control of the SoC decrease rate or the aging control (increase of γ) for the 20 consecutive days scenario, we obtain the respective aging curves of both strategies as shown in Figure (6a). Each day of discharging service resulted in an increase in the rate of γ and energy sold to the grid, as shown in Figure (6b) and Figure (6c), respectively.

The total energy sold through the SoC rate control strategy surpasses that of the γ rate control. This is because the latter prioritizes tracking the desired γ growth rate over maximizing energy discharged. In particular, SoC rate control surpasses standard discharge in total energy when it focuses utilizing the entire available discharge time. Moreover, when examining the aging rate, standard discharge emerges as the least favorable option. SoC rate control effectively mitigates aging by regulating the current rate, although it remains susceptible to random fluctuations determined by the discharge conditions. Conversely, aging rate control continuously adjusts the aging rate to achieve the required γ value by the end of the 20 cycles.

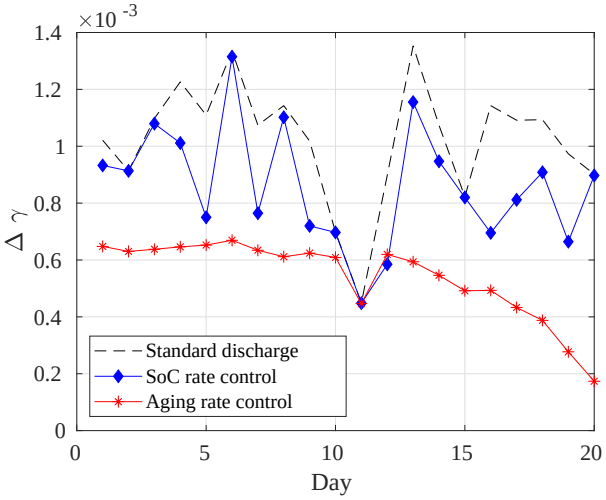
By considering 1.025 as the maximum γ rate (instead of $\gamma = 2$), the total energy sold by the battery in 20 days would be comparable to that of other strategies, as it undergoes more cycles below the maximum γ threshold. Additionally, aging rate control provides the advantage of effectively managing capacity fading. By considering the 20-day rate as a reference and considering the total aging interval ($\gamma = 2$), the battery can reach half of its maximum capacity in about 800 days with aging rate control, a level of certainty not achievable with the other strategies.

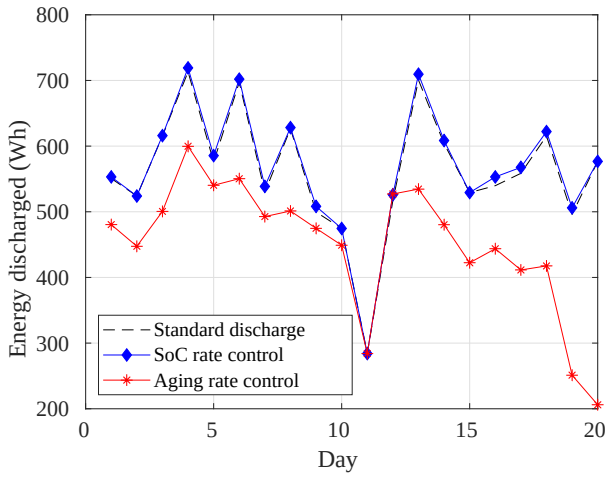
5. Conclusion

This work introduces two health-aware strategies for grid energy sales in a V2G service. The first strategy focuses on managing discharge to maintain the SoC above a specified minimum within the discharge interval, optimizing usage and



(a) Final values of $\gamma^n(t_f)$





(c) Total energy discharged to the grid

Figure 6. Results of 20 days ($N = 20$) of discharging for power selling using different strategies.

mitigating stress factors like discharge current. This strategy embodies a basic approach, based on stress factor mitigation, offering a V2G service that considers aging process. The second strategy aims to regulate the aging increase rate during the discharge event to maintain the SoH below a specified maximum within the interval discharging days. Both strategies employ adaptive control of grid demand, designed with robust techniques and error minimization. Results show the first strategy reduces the final aging, represented by the stress factors index, while increasing total energy sold at the discharging end. Conversely, the second strategy prioritizes desired degradation increase rates, potentially compromising energy sold, but it succeeds in managing the aging process. Specifically, the degradation rate regulation strategy ensures that the aging factor reaches the desired level within the specified timeframe, which proves beneficial for lifetime control. Furthermore, it still ensures the sale of energy close to the standard discharge behavior. In summary, this paper's contributions include:

- Incorporating aging effects such as capacity decrease and resistance increase into the discharge behavior of a battery model.
- Health-aware discharging approaches using degradation-rate regulation and discharge-rate regulation.

Certainly, there is significant potential in incorporating SoH and SoC estimations integrated with closed-loop HAC frameworks to effectively manage battery health. Future work involves integrating control approaches with SoC an SoH estimation approaches to validate performance when used in conjunction, particularly in embedded applications. Furthermore, the utilization of other control optimization techniques, such as Model Predictive Control (MPC), to align with different objectives, such as charging process, is encouraged. An extension of this work could involve comparing it with alternative approaches addressing the same issue. Finally, this study lays the groundwork for a charging-discharging parking service with energy selling that integrates strategies for the management of battery lifetime.

REFERENCES

Barré, A., Deguilhem, B., Grolleau, S., Gérard, M., Suard, F., & Riu, D. (2013). A review on lithium-ion battery ageing mechanisms and estimations for automotive applications. *Journal of power sources*, *241*, 680–689.

Brown, D. W., Georgoulas, G., Bole, B., Pei, H.-L., Orchard, M., Tang, L., ... Vachtsevanos, G. (2009). Prognostics enhanced reconfigurable control of electro-mechanical actuators. In *Annual conference of the phm society* (Vol. 1).

Collath, N., Tepe, B., Englberger, S., Jossen, A., & Hesse, H. (2022). Aging aware operation of lithium-ion battery

energy storage systems: A review. *Journal of Energy Storage*, *55*, 105634.

Didier, B., Thierry, P., Sébastien, M., Christian, N., Séverine, J. S. L., Bloch, D., & Séverine, J. S. L. (2021). *Li-ion batteries : development and perspectives / coordinated by didier bloch, sébastien martinet, thierry priem, ... [et al.] ; [préface de séverine jouanneau si larbi]*. Les Ulis: Science press, EDP sciences.

Félix, M. S., Martinez, J. J., & Bérenguer, C. (2023). A state-space approach for remaining useful life control. *IFAC-PapersOnLine*, *56*(2), 7728–7733.

Fricke, K., Nascimento, R., Corbetta, M., Kulkarni, C., & Viana, F. (2023). An accelerated life testing dataset for lithium-ion batteries with constant and variable loading conditions. *International Journal of Prognostics and Health Management*, *14*(2).

Kipchirchir, E., Do, M. H., Njiri, J. G., & Söffker, D. (2023). Prognostics-based adaptive control strategy for lifetime control of wind turbines. *Wind Energy Science*, *8*(4), 575–588.

Martinez, J. J., Félix, M. S., Kulkarni, C., Orchard, M., & Bérenguer, C. (2024). A novel dynamical model for diagnosis, prognosis and health-aware control of lithium-ion batteries. *Proceedings of the 12nd IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*, to appear.

Pelletier, S., Jabali, O., Laporte, G., & Veneroni, M. (2017). Battery degradation and behaviour for electric vehicles: Review and numerical analyses of several models. *Transportation Research Part B: Methodological*, *103*, 158–187.

Reniers, J. M., Mulder, G., Ober-Blöbaum, S., & Howey, D. A. (2018). Improving optimal control of grid-connected lithium-ion batteries through more accurate battery and degradation modelling. *Journal of Power Sources*, *379*, 91–102.

APPENDIX A - LQR PROBLEM

Consider a linear system represented in state-space form as:

$$\dot{x} = Ax + Bu$$

where x is the state vector, and u is the control input.

The Linear Quadratic Regulator (LQR) problem is a control strategy designed to create an optimal feedback controller for such linear systems while minimizing a quadratic cost function and stabilizing the system. The objective of the LQR problem is to minimize a quadratic cost function, defined as:

$$J = \int_0^\infty \left(x^T Q x + u^T R u\right) dt \tag{29}$$

Here, Q is a positive semidefinite weighting matrix that penalizes deviations of the state from its desired trajectory, and R is a positive definite weighting matrix that penalizes control effort or deviations of the control input from its desired values.

The optimal control law u is then determined to minimize Eq. 29 and stabilize the system. When the control law is defined as

$$u = -Kx$$

K is the optimal gain matrix found through a stability guarantee function equivalent to the Riccati Equation, which depends on the existence of a positive definite matrix P.

APPENDIX B - MODEL PARAMETERS

In battery discharging, the values of R_n and V_{oc}, as depicted in the equivalent model, vary as functions of $SoC(k)$. These variations can be expressed by the following equations:

$$R_n(SoC(k)) = M * \left(R_{min} + \frac{K_3}{SoC(k)} + \frac{K_4}{100 - SoC(k)}\right) \tag{30}$$

$$V_{oc}(SoC(k)) = M * \left(E_o - K_1 \ln\left(100 - SoC(k)\right) - \frac{K_2}{SoC(k)}\right) \tag{31}$$

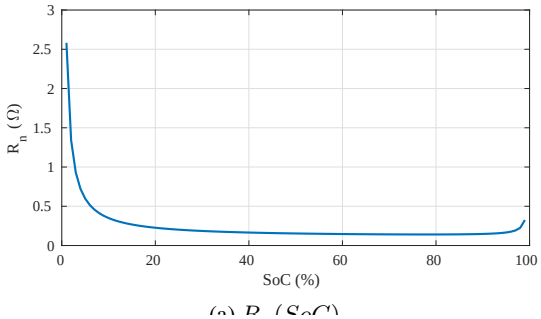
$$C_n = M * C_1 \tag{32}$$

Table 2 presents the parameters of the model used for simulation. These parameters are obtained using data from (Fricke, Nascimento, Corbetta, Kulkarni, & Viana, 2023).

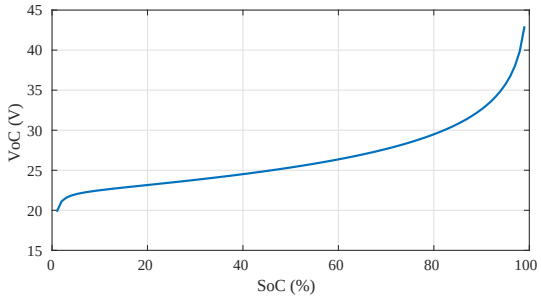
Table 2. Battery model parameters

Parameter	Value
M	10
T_s	0.02 s
a	$\exp\left(\frac{-T_s}{50}\right)$
c_3	10^{-8}
K_1	0.27
K_2	0.45
K_3	0.25
K_4	0.02

M is the number of cell in the battery pack. Obtained R_n and V_{oc} are illustrated in Figure 7.



(a) $R_n(SoC)$



(b) $V_{oc}(SoC)$

Figure 7. Nominal values of $R_n(SoC)$ and $V_{oc}(SoC)$ resulted from the considered model with mean mean parameters.

Human-Centric PHM in the Era of Industry 5.0

Parul Khanna[1], Jaya Kumari[2], and Ramin Karim[3]

[1,2,3]*Luleå University of Technology, Luleå, 97187, Sweden*
parul.khanna@ltu.se
jaya.kumari@ltu.se
ramin.karim@ltu.se

ABSTRACT

The maintenance industry is undergoing a major transformation as it embraces the shift towards Industry 5.0. The focus of Industry 5.0 is on the integration of human intelligence with advanced technologies. It emphasizes interaction and collaboration between humans and machines and aims to combine the strengths of both. The efficiency of prognostics and health management (PHM) for maintenance in industrial contexts can be enhanced by improving this human-machine interaction and collaboration. This paper investigates the human-centric aspects, with a focus on PHM systems for facilitating the enablement of Industry 5.0 in maintenance. Acknowledging human as an active participant, this study explores their integral role in designing and developing PHM systems. The data collection for this study has been based on available literature, active and passive observations, and unstructured interviews and discussions with experienced industry professionals. As a result of the analysis of collected data, this study identifies and highlights potential areas for research and exploration. The research in these areas can advance the understanding and application of human-centric PHM strategies within Industry 5.0 in maintenance contexts. This is expected to improve the resilience and sustainability aspects of the industrial ecosystem and facilitate the shift towards Industry 5.0.

Keywords— Maintenance, Prognostics and Health Management, Industry 5.0, Human-centric

1. INTRODUCTION

The industrial revolution over the years has led to systematic developments and advancements in all sectors of society. Industries have grown dramatically, starting with the mechanical revolution of the "steam engine" era (Industry 1.0), progressing towards electrical breakthroughs (Industry 2.0), developing further into the computerization and automation era (Industry 3.0), and now towards cyber-physical systems (Industry 4.0) (Leng et al., 2022). Even though Industry 4.0 is still under constant research and development, the concept of Industry 5.0 is being actively defined, discussed, and explored by academia, industry, and policymakers. It aims to extend, complement, and build on the technological advancements of its predecessor, Industry 4.0 (Raja Santhi & Muthuswamy, 2023).

The concept of Industry 5.0 is based on integrating human-centric aspects with advanced technologies for enhanced productivity and operations (Nahavandi, 2019). It emphasizes on human-centric aspects like human-machine collaboration, worker well-being, empowering workers with enhanced decision-making, and personalized/customized systems that can enhance the industry's sustainability and resilience aspects (Adel, 2022; Industry 5.0: Towards More Sustainable, Resilient and Human-Centric Industry - European Commission, n.d.). Unlike its predecessors, where the key factors were automation and technology which gave the idea of the technologies as a partial replacement for humans, Industry 5.0 seeks to integrate the strengths of both humans and machines to optimize industrial processes. In this context, it involves the integration of human intelligence, creativity, and experience, with the capabilities of advanced technologies such as AI, data analytics, IoT, etc. (Ghobakhloo et al., 2023)

PHM leverages data analytics, condition monitoring systems, digital twins, and other advanced technologies to forecast potential failures and issues before they occur, allowing proactive maintenance actions to be taken. This shift towards predictive maintenance, facilitated by PHM, helps minimize unplanned downtime, optimize maintenance costs, and improve overall asset effectiveness (Zio, 2022).

Industry 5.0 will further facilitate PHM in maintenance by focusing on human-centric aspects, which will enable maintenance workers to leverage PHM data and insights to make more informed, adaptive, and resilient maintenance decisions.

The purpose of this study is to advance the understanding and application of human-centric PHM strategies within Industry

5.0 maintenance contexts and make the industrial ecosystem more resilient and sustainable.

The objectives of this study are:

1. To investigate the human-centric aspects of Industry 5.0 in maintenance, focusing on PHM systems.
2. To identify the potential areas for research and exploration for advancing the understanding and application of human-centric maintenance strategies within the context of Industry 5.0.

2. METHODOLOGY

This study was conducted using a mixed-method research approach to investigate the human-centric aspects, with a specific focus on PHM systems for facilitating the enablement of Industry 5.0 in maintenance. The methodology followed consisted of the following components:

- Literature Review: A review of existing literature was conducted on Industry 4.0 and 5.0, human-centric aspects, PHM systems, and maintenance efficiency to establish a theoretical foundation for the study.
- Observations: To understand the practical implementations and dynamics of human-machine interactions in industrial maintenance, active and passive observations were made. This included observing industrial professionals utilizing PHM systems in real-world settings to see how they interact with these systems during their daily operations. Observing demonstrations of maintenance systems in laboratory settings during lab visits, which included a diverse audience ranging from students to industry professionals.
- Unstructured Interviews and discussions: Conducted unstructured interviews and discussions with asset managers from companies serving as knowledge partners for rail vehicles and public transport agencies in Sweden. These interviews/discussions were conducted during workshops, seminars, and regular meetings. They provided strategic insights into the adoption and challenges of PHM systems. They additionally provided us with valuable insights from maintenance worker's perspective enabling us to gather firsthand information on their experiences, challenges, and perspectives regarding the integration of human aspects into PHM systems.
- Data Analysis: Qualitative analysis of data collected from literature surveys, observations, and interviews to identify key themes, patterns, and challenges associated with human-centric PHM strategies in Industry 5.0 maintenance contexts.

3. LITERATURE REVIEW

A review of existing literature was conducted on Industry 4.0, and Industry 5.0 in connection with maintenance and PHM systems to establish a theoretical foundation for the study. Since human-centricity is a key factor in moving towards Industry 5.0, we conducted a literature review on human-centric aspects in maintenance and PHM. These considered reviews were from the period 2014 - 2024. To help understand the Industrial Revolution journey, works focusing on the revolutions thus far were also considered. These works dates from 1956 till the present. Key search terms included "Industry 4.0 AND PHM", "Industry 4.0 AND maintenance", "Industry 5.0 AND PHM", "Industry 5.0 AND maintenance", "Human-centric AND PHM", and "Human-centric AND maintenance". In total, the study was conducted with 26 relevant works of literature.

3.1. Industrial Revolutions Leading to Industry 5.0

The Industrial Revolutions, over the years, have helped in shaping the current industrial landscape. Starting in the late 18th century, the first industrial revolution was enabled by the mechanical revolution and the usage of steam power resulting in faster production processes (Martinelli et al., 2021). Subsequently came the 2nd Industrial Revolution, which focused on the electrical revolution for mass production techniques and implementing assembly lines. Industry 3.0 brought the digital revolution with IT and automation transformations which resulted in significant technological advancements and societal changes.

The current Industrial Revolution i.e. Industry 4.0 saw an increase in Cyber-Physical Systems, IoT and AI which focused on the integration of technology with physical assets. The focus has been primarily on the technological aspects, with limited attention paid to the human and social factors within organizations (Moraes et al., 2023). Extending this is the concept of Industrial 5.0, which works alongside the technological advancements till Industry 4.0 but puts humans in the centre of it. Figure **1** shows an advancement of industrial processes from an abstract level.

The revolutionary journey from the first industrial revolution to Industry 4.0, which focused on a technology-driven approach emphasizing digitalization and advanced technologies like digital twins, AI and cybersecurity, laid the foundation for Industry 5.0 (Nagano, 2019). Industry 5.0, as introduced by the European Commission, (*Industry 5.0 - European Commission*, n.d.) represents a shift towards a user-centric and value-driven approach, emphasizing the crucial role of humans in the industrial process and promoting principles of social well-being, sustainability, and human-machine collaboration (Beaudreau, 2018; Verma et al., 2022).

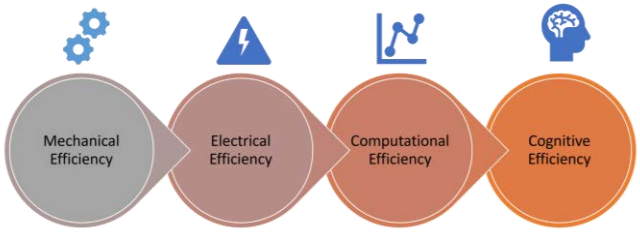


Figure 1: Advancing of Industrial Processes

Industrial revolutions have had significant impacts on maintenance practices, influencing how industries manage and upkeep their machinery and equipment. Table **1** (adapted by (Poór et al., 2019) shows the relationship between industrial revolutions, their enablers and key maintenance facilitators(Coleman, 1956). It is important to note that the mentioned enablers and maintenance facilitators extend and complement their predecessor's enablers and maintenance facilitators respectively.

Table 1: Industrial Revolutions- enablers and key maintenance facilitators

Industrial Revolution	Enablers	Key maintenance facilitators
Industry 1.0	Mechanical Revolution, Steam Power	Visual Inspection
Industry 2.0	Electrical Revolution, Mass production	Instrumental/Tool Inspection, Preventive Maintenance
Industry 3.0	Digital Revolution, Automation	Sensors, CMMS, Predictive Maintenance
Industry 4.0	Cyber-Physical Systems, IoT, AI, ML	Data Analytics (Predictive Analytics), Digital Twins, Condition-Based Maintenance
Industry 5.0	Human-Machine Collaboration, AI, ML	HSI, Advanced Predictive Analytics, AR/VR, Blockchain

3.2. Industry 5.0 and its implications for the maintenance industry

According to the European Commission (*Industry 5.0 - European Commission*, n.d.; *Industry 5.0: Towards More Sustainable, Resilient and Human-Centric Industry - European Commission*, n.d.), implementing Industry 5.0 means placing the well-being of humans at the centre of the industrial processes. It encourages the usage of advanced technologies to focus beyond productivity and efficiency and emphasizes the well-being of the human workforce while considering the planet's resource constraints. It builds on the existing industrial revolution i.e. Industry 4.0 and compliments it while focusing on three key factors, Human-centricity, Sustainability and Resilience (Figure 2).

Industry 5.0 emphasizes the collaboration between humans and machines, focusing on enhancing human creativity and well-being while leveraging advanced technologies like big data analytics, IoT, collaborative robots (cobots), Blockchain, digital twins, and future 6G systems (Adel, 2022; *Industry 5.0 - European Commission*, n.d.).

The impact of Industry 5.0 on the maintenance industry is profound. It implies the usage of modern advanced technologies with a human-centric approach to sustainable and resilient maintenance processes. It involves data-driven decision-making that addresses potential maintenance faults before they lead to breakdowns, optimising operational efficiency, reducing downtime, keeping customers satisfied, and contributing to sustainability efforts by focusing on repair and recycling rather than replacement (Psarommatis et al., 2023).

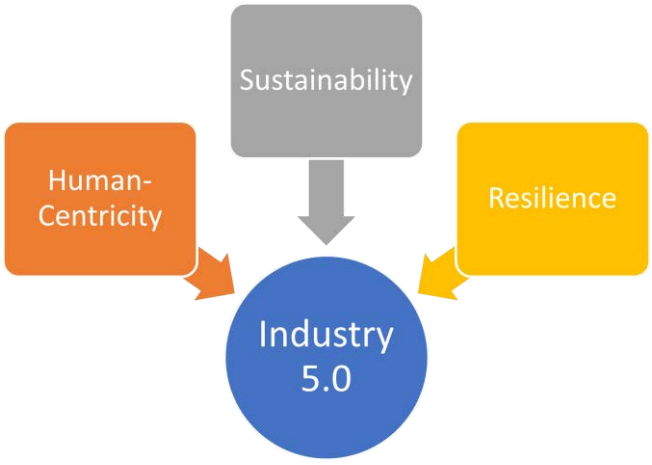


Figure 2: 3 pillars of industry 5.0

3.3. Prognostics and Health Management (PHM) Systems and Industry 5.0

PHM systems are designed to monitor the health of industrial assets. In the context of Industry 5.0, PHM systems can be seen as a cornerstone for the integration of modern advanced technologies like big data analytics, IoT, and A) with human-centric approaches within industrial maintenance practices. It enhances the efficiency and effectiveness of maintenance

practices, operational efficiency, and sustainability of industrial operations. (Biggio & Kastanis, 2020) (Adel, 2022).

Industry 5.0 places human well-being at the centre of industrial processes. PHM systems support this approach by empowering maintenance workers with real-time insights and decision-support tools. By providing workers with data-driven insights about equipment health and performance, PHM systems help integrate human intelligence with these insights to make informed decisions, optimize maintenance activities, and ensure the safety and well-being of workers in industrial environments (Kumar et al., 2023).

PHM systems also play an important role in ensuring the sustainability and resilience of industrial maintenance processes. By forecasting asset failures and scheduling maintenance actions at optimum times, these systems help to minimize waste and reduce the environmental impact of manufacturing operations contributing to the sustainability factor(Ghobakhloo et al., 2024). Additionally, PHM systems contribute to the implementation of smart factories, which are a key aspect of Industry 4.0 and Industry 5.0, by providing real-time insights into the health of industrial assets and enabling more efficient and effective maintenance strategies which contribute to enhancing the resilience of the industrial assets (Kumar et al., 2023).

Predictive maintenance (PdM) which is facilitated by PHM systems monitors the health of industrial assets, predicts potential failures, and optimizes maintenance schedules based on the predicted future state of equipment components. By adopting a human-centric approach to PdM within the Industry 5.0 framework, organizations can enhance decision-making processes, increase trust between decision-makers and predictive models, allocate resources effectively, and improve overall maintenance effectiveness(van Oudenhoven et al., 2023).

Therefore, the integration of human-centric maintenance practices within the principles of Industry 5.0 enables proactive management of maintenance needs, reduces costs, enhances operational efficiency, ensures equipment reliability, and contributes to sustainable maintenance practices by focusing on repair and recycling rather than replacement. This connection highlights the importance of predictive maintenance which is facilitated by PHM systems as a key enabler of Industry 5.0's vision for smarter, more efficient, effective, and human-centred maintenance processes.

3.4. Human involvement in designing and developing PHM systems

In the era of advanced technologies, a human-centric approach to developing PHM systems for industrial maintenance is not just desirable but essential. A human-centric PHM system empowers users with intuitive interfaces, actionable insights, and decision support tools to optimize maintenance strategies ultimately leading to more efficient and effective maintenance activities.

Involving domain experts and their insights into industrial processes, especially maintenance activities, aids in focusing on critical aspects that are prone to failure. Domain experts collaborate with developers early on to define system requirements tailored to operational contexts, and technician knowledge levels (Toothman et al., 2023). Humans can also consider factors like production load, environmental conditions, and other maintenance activities that may influence asset health, which algorithms might overlook (McDonnell et al., 2018).

Humans also play a critical role in selecting the relevant data points for training and monitoring equipment health, ensuring the quality of data used in PHM systems and interpreting system outputs (Siew et al., 2020). Additionally, Usability Engineering and Usability Requirement Analysis are critical areas where specialists ensure that PHM systems meet the maintenance personnel's needs and requirements in a user-friendly manner.

Usability Engineering aspects emphasize visually presenting clear explanations, minimizing cognitive load to enhance usability and ensure effective decision-making for maintenance personnel (McDonnell et al., 2014). This human-centred approach is essential to optimize the functionality and user-friendliness of PHM systems, making them more accessible and efficient for operators.

Usability Requirement Analysis, on the other hand, focuses on identifying and documenting the usability needs and objectives of the system. This involves gathering and analysing user requirements related to usability, accessibility, and user experience, providing a framework for designing and evaluating the user interface.

Customized PHM dashboards can prioritize relevant data points for specific tasks and assets, aiding in quicker issue identification and efficient maintenance actions. Alerts and notifications can also be tailored according to their criticality reducing information overload and ensuring timely response to critical issues(McDonnell et al., 2014).

Another interesting area is to investigate the legitimacy aspect of PHM systems for a necessary understanding of why such predictions were made, fostering trust in the system and the recommendations made by it. It will encourage confident decision-making by the technicians. This comes under the umbrella of Explainable AI for trust and continuous improvement. It enables debugging in case of incorrect recommendations, and human-in-the-loop learning for continuous improvement of algorithms. By understanding the reasoning behind predictions, humans can detect and address these issues, leading to improved accuracy and reliability of the PHM system (Amin et al., 2022; Nor et al., 2021).

4. RESULTS

This research highlights some key insights while integrating human-centric aspects with PHM systems within the context of Industry 5.0 and industrial maintenance.

Following are the findings for the objectives of this study:

Objective 1: To investigate the human-centric aspects of Industry 5.0 in maintenance, focusing on PHM systems.

Industry 5.0 places the well-being of humans at the centre of industrial processes. It emphasizes the collaboration between humans and machines. PHM systems play an important role in this approach, empowering maintenance workers with valuable insights and decision-support tools. These systems when integrated with human intelligence and data-driven insights can optimize maintenance activities and ensure worker safety and well-being.

Integrating human aspects into PHM systems involves real-time collaboration between human operators and machines and the incorporation of human-in-the-loop mechanisms. These possibilities aim to enhance the usability, acceptance, and integration of PHM systems within industrial work environments.

Objective 2: To identify potential areas for research and exploration for advancing the understanding and application of human-centric maintenance strategies within the context of Industry 5.0.

The identified areas for further exploration within the context of Industry 5.0 and industrial maintenance especially PHM systems include human-system interaction, Explainable AI, Usability Requirement Analysis and Usability Engineering. These areas highlight the need for exploring the dynamics of human-machine collaboration and identifying strategies to optimize human-system interactions for improved maintenance activities and enhanced decision-making. Observations and discussions with maintenance professionals have highlighted the critical role of seamless human-system interaction in enhancing operational efficiency. Additionally, future research will delve into Explainable AI, to enable maintenance personnel to understand the reasoning behind AI-generated predictions and recommendations, fostering trust in the system and facilitating human-in-the-loop processes for continuous improvement. Insights from unstructured interviews and discussions emphasized the importance of transparency in AI systems for maintenance workers. Furthermore, Usability Requirement Analysis will play a pivotal role in identifying and prioritizing user needs and preferences in the context of PHM systems. This area benefits significantly from feedback gathered through interviews and discussions with asset managers. Usability Engineering will play a crucial role in designing user-centric interfaces and interactions for PHM systems in the Industry 5.0 context to enhance the usability and user experience for maintenance personnel. Laboratory observations and real-world use cases have provided valuable insights into the need for creating more effective and user-friendly maintenance systems. Figure 3 shows the key insights of Industry 5.0 within the industrial maintenance context.

The transition to Industry 5.0 introduces several innovative challenges as compared to conventional maintenance activities. These include the integration of advanced technologies which require new skills and training for the users/workers. The focus shifts towards creating more responsive, robust, and resilient maintenance systems.

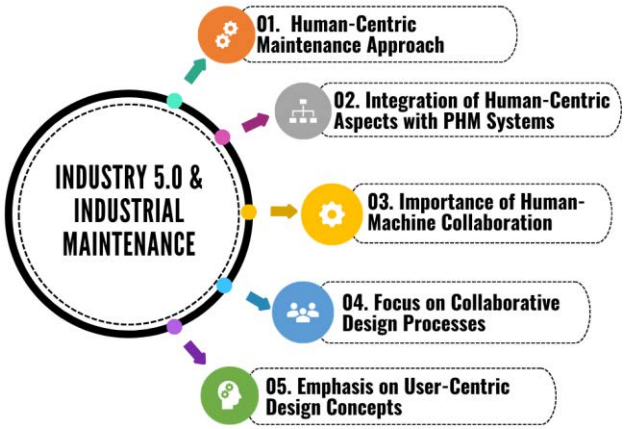


Figure 3: Key insights of Industry 5.0 within the industrial maintenance context

5. CONCLUSION

The maintenance industry is constantly developing and with the latest industrial revolution i.e. Industry 5.0, the shift is towards a human-centric maintenance approach. The research on industrial maintenance especially the PHM systems within the context of Industry 5.0 is important. It emphasizes the significance of the role of human involvement in optimising maintenance practices. The integration of human-centric aspects with PHM systems also enables industries to improve asset reliability and enhance operational efficiency. Furthermore, the findings emphasize the importance of human-machine collaboration, and data-driven decision-making in realizing the full benefits of Industry 5.0 in maintenance operations. While interviews primarily focused on railway experts, the findings can be extended to various industrial settings, indicating broader applicability. This research highlights the need to focus on collaborative design processes and user-centred approaches to ensure effective human-machine interactions, which are essential for the successful implementation of maintenance practices within the context of Industry 5.0.

The future research will be built on the findings and insights from this research. Future work will focus on achieving a more seamless and efficient human-machine interface, considering human-centric aspects during system design and implementation. As a result, maintenance activities will eventually be more effective and efficient. An important area of research is quantifying the effectiveness and efficiency of such human-centric PHM systems.

ACKNOWLEDGEMENT

We gratefully acknowledge the European Commission for its support of the Marie Sklodowska Curie program through the H2020 ETN MOIRA project (GA 955681). We also acknowledge the valuable support and resources provided by the eMaintenanceLAB in conducting this research.

NOMENCLATURE

AI	Artificial Intelligence
IoT	Internet of Things
PHM	Prognostics and Health Management
XAI	Explainable Artificial Intelligence
CMMS	Computerized Maintenance Management System
HSI	Human-System Interaction
AR	Augmented Reality
VR	Virtual Reality

REFERENCES

Adel, A. (2022). Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. *Journal of Cloud Computing 2022 11:1*, *11*(1), 1–15. https://doi.org/10.1186/S13677-022-00314-5

Amin, O., Brown, B., Stephen, B., & McArthur, S. (2022). A Case-study Led Investigation of Explainable AI (XAI) to Support Deployment of Prognostics in the industry. *PHM Society European Conference*, *7*(1), 9–20. https://doi.org/10.36001/PHME.2022.V7I1.3336

Beaudreau, B. C. (2018). A Pull–Push Theory of Industrial Revolutions. *International Advances in Economic Research*, *29*(4), 303–317. https://doi.org/10.1007/S11294-023-09883-W

Biggio, L., & Kastanis, I. (2020). Prognostics and Health Management of Industrial Assets: Current Progress and Road Ahead. *Frontiers in Artificial Intelligence*, *3*, 578613. https://doi.org/10.3389/FRAI.2020.578613/BIBTEX

Coleman, D. C. (1956). Industrial Growth and Industrial Revolutions. In *New Series* (Vol. 23, Issue 89).

Ghobakhloo, M., Hannan, ·, Mahdiraji, A., Iranmanesh, M., & Vahid Jafari-Sadeghi, ·. (2024). From Industry 4.0 Digital Manufacturing to Industry 5.0 Digital Society: a Roadmap Toward Human-Centric, Sustainable, and Resilient Production. *Information Systems Frontiers 2024*, 1–33. https://doi.org/10.1007/S10796-024-10476-Z

Ghobakhloo, M., Iranmanesh, M., Tseng, M. L., Grybauskas, A., Stefanini, A., & Amran, A. (2023). Behind the definition of Industry 5.0: a systematic review of technologies, principles, components, and values. *Journal of Industrial and Production Engineering*, *40*(6), 432–447. https://doi.org/10.1080/21681015.2023.2216701

Industry 5.0 - European Commission. (n.d.). Retrieved March 23, 2024, from https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en

Industry 5.0: Towards more sustainable, resilient and human-centric industry - European Commission. (n.d.). Retrieved March 24, 2024, from https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/industry-50-towards-more-sustainable-resilient-and-human-centric-industry-2021-01-07_en

Kamal, A., Nor, M., Rao Pedapati, S., & Muhammad, M. (n.d.). *Explainable AI (XAI) for PHM of Industrial Asset: A State-of-The-Art, PRISMA-Compliant Systematic Review*.

Kumar, P., Raouf, I., & Kim, H. S. (2023). Review on prognostics and health management in smart factory: From conventional to deep learning perspectives. *Engineering Applications of Artificial Intelligence*, *126*, 107126. https://doi.org/10.1016/J.ENGAPPAI.2023.107126

Leng, J., Sha, W., Wang, B., Zheng, P., Zhuang, C., Liu, Q., Wuest, T., Mourtzis, D., & Wang, L. (2022). Industry 5.0: Prospect and retrospect. *Journal of Manufacturing Systems*, *65*, 279–295. https://doi.org/10.1016/J.JMSY.2022.09.017

Martinelli, E. M., Farioli, M. C., & Tunisini, A. (2021). New companies' DNA: the heritage of the past industrial revolutions in digital transformation. *Journal of Management and Governance*, *25*(4), 1079–1106. https://doi.org/10.1007/S10997-020-09539-5

McDonnell, D., Balfe, N., Al-Dahidi, S., & O'Donnell, G. E. (2014). Designing for Human-Centred Decision Support Systems in PHM. *PHM Society European Conference*, *2*(1). https://doi.org/10.36001/PHME.2014.V2I1.1558

McDonnell, D., Balfe, N., Pratto, L., & O'Donnell, G. E. (2018). Predicting the unpredictable: Consideration of human and organisational factors in maintenance prognostics. *Journal of Loss Prevention in the Process Industries*, *54*, 131–145. https://doi.org/10.1016/J.JLP.2018.03.008

Moraes, A., Carvalho, A. M., & Sampaio, P. (2023). Lean and Industry 4.0: A Review of the Relationship, Its Limitations, and the Path Ahead with Industry 5.0. *Machines*, *11*(4). https://doi.org/10.3390/MACHINES11040443

Nagano, A. (2019). Thinking about industrial revolutions in systems theory - Moving towards the fourth industrial

revolution. *ACM International Conference Proceeding Series*, *Part F148155*(2), 470–471. https://doi.org/10.1145/3326365.3326429

Nahavandi, S. (2019). Industry 5.0—A Human-Centric Solution. *Sustainability 2019, Vol. 11, Page 4371*, *11*(16), 4371. https://doi.org/10.3390/SU11164371

Nor, A. K. M., Pedapati, S. R., Muhammad, M., & Leiva, V. (2021). Overview of Explainable Artificial Intelligence for Prognostic and Health Management of Industrial Assets Based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses. *Sensors (Basel, Switzerland)*, *21*(23). https://doi.org/10.3390/S21238020

Poór, P., Ženíšek, D., & Basl, J. (n.d.). *Historical Overview of Maintenance Management Strategies: Development from Breakdown Maintenance to Predictive Maintenance in Accordance with Four Industrial Revolutions*.

Psarommatis, F., May, G., & Azamfirei, V. (2023). Envisioning maintenance 5.0: Insights from a systematic literature review of Industry 4.0 and a proposed framework. *Journal of Manufacturing Systems*, *68*, 376–399. https://doi.org/10.1016/J.JMSY.2023.04.009

Raja Santhi, A., & Muthuswamy, P. (2023). Industry 5.0 or industry 4.0S? Introduction to industry 4.0 and a peek into the prospective industry 5.0 technologies. *International Journal on Interactive Design and Manufacturing (IJIDeM) 2023 17:2*, *17*(2), 947–979. https://doi.org/10.1007/S12008-023-01217-8

Siew, C. Y., Chang, M. M. L., Ong, S. K., & Nee, A. Y. C. (2020). Human-oriented maintenance and disassembly in sustainable manufacturing. *Computers & Industrial Engineering*, *150*, 106903. https://doi.org/10.1016/J.CIE.2020.106903

Toothman, M., Braun, B., Bury, S. J., Moyne, J., Tilbury, D. M., Ye, Y., & Barton, K. (2023). Overcoming Challenges Associated with Developing Industrial Prognostics and Health Management Solutions. *Sensors 2023, Vol. 23, Page 4009*, *23*(8), 4009. https://doi.org/10.3390/S23084009

van Oudenhoven, B., Van de Calseyde, P., Basten, R., & Demerouti, E. (2023). Predictive maintenance for industry 5.0: behavioural inquiries from a work system perspective. *International Journal of Production Research*, *61*(22), 7846–7865. https://doi.org/10.1080/00207543.2022.2154403

Verma, A., Bhattacharya, P., Madhani, N., Trivedi, C., Bhushan, B., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Blockchain for Industry 5.0: Vision, Opportunities, Key Enablers, and Future Directions. *IEEE Access*, *10*, 69160–69199. https://doi.org/10.1109/ACCESS.2022.3186892

Zio, E. (2022). Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*, 108119. https://doi.org/10.1016/J.RESS.2021.108119

Hybrid AI-Subject Matter Expert Solution for Evaluating the Health Index of Oil Distribution Transformers

Augustin Cathignol[1], Victor Thuillie-Demont[2], Ludovica Baldi[3], Laurent Micheau[4], Jean-Pierre Petitpretre[5], Amelle Ouberehil[6]

[1,2,3] *Schneider Electric, 160 Avenue des Martyrs, 38000 Grenoble, France*
augustin.cathignol@se.com, victor.thuilliedemont@se.com, ludovica.baldi@se.com

[4,5,6] *Transfo Services, Rue Jacques Lieutaud, 13200, France*
laurent.micheau@se.com, jean-pierre.petitpretre@se.com, amelle.ouberehil@se.com

ABSTRACT

Reliability of oil distribution transformers is paramount, ensuring a stable flow of electricity and shielding from potential fire hazards. The internal insulation system of these transformers utilizes a combination of oil and paper. As the oil circulates through the active part of the system, it collects gaseous and physical traces of existing or past defects or degradations, providing a holistic view of the transformer's health, and allowing for early detection of problems and predictive maintenance. While various and mainly data-driven methods have been developed to calculate a transformer health index from oil samples, they lack accuracy due to limited data. This paper proposes a novel hybrid approach that leverages both Artificial Intelligence and Subject Matter Expertise to enhance the health estimation of oil distribution transformers. Our methodology utilizes a substantial dataset exceeding 65,600 analyzed oil samples, coupled with the valuable knowledge of domain experts. This combined approach achieves an accuracy exceeding 95%, suitable for real-world industrial applications. Furthermore, we introduce a risk management feature that strengthens the ability to identify transformers at high risk of failure. Notably, the health index estimation is implemented as a semi-automatic process, retaining the "expert in the loop" principle for managing critical and ambiguous cases.

1. INTRODUCTION AND PROBLEM STATEMENT

Distribution transformers convert high-voltage electricity from transmission lines into usable power for homes, businesses, and industries. Their reliability is paramount, ensuring a stable and continuous flow of electricity, shielding us from power outages, and potential fire hazards. Regular inspections, advanced fault detection systems, condition monitoring, and proper maintenance are crucial for those transformers. The internal insulation system of a transformer is provided by both oil and paper. As the oil circulates through the active part of the system, it collects gaseous and physical traces of existing or past defects or degradations. Therefore, it provides a holistic view of the transformer's health, allowing for early detection of problems and enabling predictive maintenance. In such a process, oil samples are extracted and analyzed in the laboratory regarding their physical and chemical properties (dielectric strength, acidity, humidity, color, and dissolved gas concentration). Sample extraction and analysis are done on a regular basis that can range from a several months to year periodicity. The results are interpreted by experienced subject matter experts who attribute a Health Index (HI) to the transformer and guide maintenance actions that could be required. The huge number of oil distribution transformers currently in operation and the limited number of experienced subject matter experts available to estimate the HI of these devices, motivates to support them. Several methods were developed to automatically compute the HI. A review of HI automatic assessment techniques for distribution and power transformers was proposed by Quynh T. Tran, Kevin Davies, Leon Roose, Puthawat Wiriyakitikun, Jaktupong Janjampop, Eleonora Riva Sanseverino and Gaetano Zizzo (2020). Some of the techniques rely on on-line data, which is not the scope of this study. Most of the techniques that rely on off-line data are fully data driven. The HI estimations done by experts not only rely on standardized combinatory calculations, but also reflect the human expertise in interpreting results. Consequently, it is difficult to translate them into mathematical formulas and several studies implemented fuzzy logic approaches, as proposed by Ahmed E. B. Abu-Elanien M.M.A. Salama, and M. Ibrahim (2012). Whatever the data-driven algorithm used, these methods are poorly

explainable, which reduces their capacity to be adopted in real operations. Only some of them rely on hybrid approach combining expert knowledge, either for feature selection, as proposed by Khalil Ibrahim, R.M. Sharkawy, H.K. Temraz and M.M.A. Salama (2016), or for uncertainty management, using Bayesian modeling, as proposed by P. Sarajcev, D. Jakus, J. Vasilj, M. Nikolic (2018). Finally, while these approaches offer intriguing avenues, they rely on very limited data set, with fewer than 100 samples (Ahmed E. B. et al. (2012), P. Sarajcev et al (2018), Atefeh Dehghani Ashkezari et al. (2012), Jahanzaib Javid et al. (2021), Ahmed E. B. Abu-Elanien et al. (2011)). These restricted datasets are unlikely to encompass the exhaustive spectrum of parameter combinations, rendering them poorly fit for real-world implementation, especially considering the safety concerns.

Now, the motivation of this work is precisely to build a prognostics and health management solution that is suitable for industrial usage, meaning performant enough, resilient enough, and preserving safety in any case. The core of this work is an original hybrid approach relying on both Artificial Intelligence (AI) and Subject Matter Expertise (SME), making the most of AI and human expertise. Practically, this approach is supported by more than 60 000 oil samples that were analyzed and from which experts provided Health Index estimations. Indeed, the idea is to train a Machine Learning (ML) algorithm to estimate HI from oil samples analyses results (Figure 1). The performance criterion that is pursued is the global accuracy of the health estimation, with a special focus on the ability of detecting transformers at risk, for obvious security reasons. The solution is also expected to be explainable and to be suitable with an "keeping expert in the loop" approach. Indeed, the target solution is not a fully automated solution, but rather a mostly automated solution that will keep experts in the loop for managing the most ambiguous and critical cases.

The paper describes the health estimation global solution, starting with the data science steps, from data collection, data cleaning, outliers' detection, imputation for managing missing data, up to the model selection and validation. It also emphasizes the way subject matter expertise was combined with AI techniques. It describes the risk management method that was introduced to minimize the risk of failing to detect at risk transformers. Finally, it practically describes the global health evaluation process in an industrial context with a "keeping expert in the loop" approach that was mentioned above.

2. HYBRID HEALTH ESTIMATION METHOD

2.1. Introduction

The method that is used is said hybrid approach as it is based on both machine learning and subject matter expertise. The machine learning pillar is a standard approach from a data science point of view, supported by subject matter expertise at all stages (feature engineering, outliers' detection, missing value management, result validation). The expertise is also explicitly integrated in a rules-based approach that complements the machine learning approach to refine health estimation. The current section covers the following technical data science steps: data collection, data cleaning, outliers' detection and validation, missing values management, models benchmarking, including oversampling and/or subsampling methods, and rule-based classification.

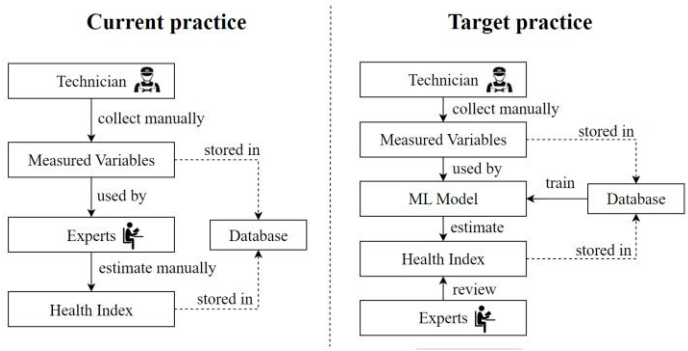


Figure 1: Current and target practice.

2.2. Data collection

This study focuses on distribution transformers which power is less than 3150 kVA and using mineral oil. Data from oil analyses over the past 10 years were used, representing approximately 65,600 analyses for 40,000 distinct transformers (as some of them have been analyzed several times all along their lifecycle). The predictive variables identified in the data are the levels of dissolved gases, color, acidity and humidity, as the dielectric strength. The target of our study is the Health Index (HI), which was estimated by experienced subject matter experts, based on the oil analysis result. Predictive variables and target are given in Figure 2.

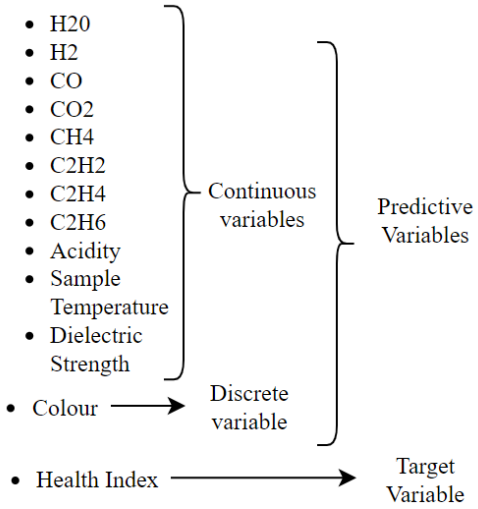


Figure 2: Predictive variables and target.

The distribution of the Health Index is intrinsically highly unbalanced: for most of the analyses the Health Index is evaluated by experts at 1 (around 90%), followed by a smaller number of analyses with a Health Index of 2 (around 9%), and finally an even smaller number of analyses with a Health Index of 3 (around 1%). A Health Index equal to 1 means that the transformer is perfectly healthy, a Health Index of 2 is an intermediate status showing some anomalies, non-serious at this stage but requiring some surveillance, and a Health Index of 3 means critical anomalies that require immediate maintenance actions.

2.3. Outliers' detection

The data cover more than 10 years of oil analyses carried out by technical experts. As the analysis process is manual, it may be possible to have some errors in the data. In this study potential errors were tracked using an anomaly detection approach. Then, each anomaly was reviewed, confirmed or not to be an error, by technical experts. The anomaly detection used in this study relies on a statistical test: the Hotelling's T-squared test. Hotelling's test is a multivariate statistical test used to determine if there are significant differences between the means of two groups in a multivariate space. In anomaly detection, it allows us to compare the mean vector and covariance matrix of a single data point (or a small group of data points) to those of a reference group. If the data point falls outside a certain threshold based on the T-squared statistic, it's flagged as an anomaly. When using Hotelling's T-squared test, it's important to ensure that the data follow a multivariate normal distribution, as it is an assumption it relies on. In the present case, dissolved gas concentrations follow an exponential distribution, so a logarithmic transformation was needed before running the test. Using a 99% confidence level for setting the T-squared threshold, 125 analyses (out of a total of 65,600 analyses, meaning 0.2%) were identified with potential errors. As atypical does not mean error, it was necessary to have those analyses reviewed by technical experts to confirm whether or not the presence of errors. In the end only 17 analyses were confirmed with errors and removed from the dataset. For the other analyses initially flagged the level of some dissolved gases was exceptionally high but perfectly credible.

2.4. Models benchmarking

To choose the machine learning algorithm that will be the basis of our hybrid approach, we perform a model screening, by doing a classification test with nine different algorithms, on all analyses for which no data are missing. The imputation method for managing missing data is studied later, after finetuning the hyperparameters of each algorithm. According to the results shown in Table 1 and obtained with a 12,000 analyses validation dataset, it appears that the most performant algorithm is the Random Forest Classifier. Accuracy is not the only performance criterion that we want to meet. Indeed, interpreting the model and validating, to some level, its consistency with the subject matter expert's way of working is another relevant criterion. A Random Forest algorithm doesn't allow to clearly identify the reasons behind a given prediction, as the final output is a combination of many decision trees, making it difficult to pinpoint the logic for each prediction. However, Random Forest algorithms usually embed feature importance techniques that show how much a specific feature contributes to the overall mode. Such a technique was used, and it could be verified that the top 3 most influencing features, Hydrogen (H2), dielectric strength and acetylene (C2H2), are consistent with the subject matter expertise, which historically allows to estimate the HI. Indeed, Hydrogen (H2) is the gas produced by most technical faults, so an analysis of oil sample done by experts always starts with this gas. The dielectric strength (also called rigidity) provide experts with a good indication of the water present in the transformer oil and therefore of the risk during operation: a too low dielectric strength induces a risk of flashover. Even in the field, this parameter is checked after certain maintenance operations, before restarting the equipment. Finally, acetylene (C2H2) is synonymous for experts with an electric arc, and therefore a major electrical fault. This consistency between algorithm feature importance and subject matter expertise gives trust in health estimation algorithm.

Table 1: Classifiers' performances obtained on the validation dataset.

Classifier	Accuracy	AUC	Recall	Precision
Random Forest	0.96	0.93	0.96	0.95
Extreme Gradient Boosting	0.96	0.92	0.96	0.95
Light Gradient Boosting	0.96	0.93	0.96	0.95
Gradient Boosting	0.96	0.93	0.96	0.95
Extra Tree	0.95	0.92	0.95	0.95
Ada Boost	0.95	0.85	0.95	0.95
Logistic Regression	0.94	0.82	0.94	0.93
Decision Tree	0.94	0.79	0.94	0.94
K Neighbors	0.93	0.67	0.93	0.90

2.5. Imputation of missing values

In the dataset used for the study (representing 65,600 analyses), the data corresponding to the dissolved gas concentrations are complete for all analyses (no missing value for any of the dissolved gases). Concerning the other

data, a value is missing for less than 10% of the analyses. Different approaches to impute missing values were tested. The goal of this step is to make the most of all analyses, even those for which a value is missing. First, the simplest imputation method was chosen as a reference: imputation by the mean. This method consists in replacing the missing values for a given feature by the average value of the feature itself. Then, the data imputation was tested using two other methods that follow a similar approach: iterative imputation. Rather than simply replacing missing values with point estimates, iterative imputation makes multiple passes over the data, using the observed values to estimate and fill in the missing values, and then repeating this process several times to improve the estimates. This allows for data variability and relationships between variables, providing more robust estimates. Iterative imputation methods may include statistical models or machine learning techniques to estimate missing values. wo variants were tested. The first one is a standard version of iterative imputation, that includes a linear regression to estimate missing values. This first variant is promising as there are significant correlations between some of the variables, as can be seen in Figure 3. Indeed, for each analysis, each input parameter can be quite well estimated thanks to the others. The second variant, also called Miss Forest, uses Random Forest models to predict and fill in missing data. It builds separate models for variables with missing data, using the available data to make accurate predictions. This method is effective for handling both continuous and categorical variables.

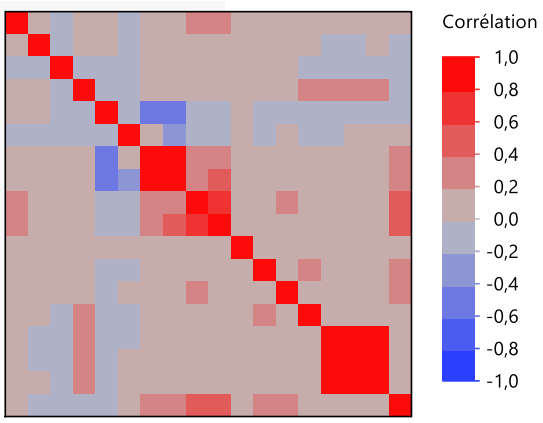


Figure 3: Correlation between parameters. Each row and each column are input variables (features).

Here we compare the output results of the Random Forest classifier using the three different imputers, on the analyses for which at least one value is missing. They all lead to the same accuracy, and according to the results shown in Table 2, it appears that the iterative imputation using a linear regression shows the best results in terms of precision for HI=1 class, recall for HI=3 class and proportion the falsely predicted HI 1 instead of 3. Therefore, the classic iterative imputer was chosen as imputation method.

Table 2: Imputation using iterative imputer, simple (mean) imputation, and Miss Forest.

Imputation method	Precision HI=1	Recall HI=3	Proportion of HI 3 diagnosed as HI 1
Iterative Imputer (Linear Regression)	**0.97**	**0.63**	**0.17**
Simple Imputation (Mean)	0.96	0.57	0.17
Miss Forest	0.96	0.57	0.38

2.6. Subject matter expert – Rule based Classification

2.6.1. Compliance with normative values

After having apply technical rules that must be respected to remain in compliance with the normative values, the expert predicted a Health Index. Concretely, these rules allow to frame the result. It is thus impossible in our context to highlight for a given analysis whose values would have exceeded the thresholds set by the normative standards.

2.6.2. Rules based on evolution through time

In most cases, a single analysis allows to set the HI of a transformer, but in some ambiguous cases, experts do use the past analyses of the same transformer (up to two additional past analyses) to refine their diagnostic. By combining last analysis results with the evolution in the dissolved gas concentrations between successive analyses, experts set the final HI. Such an approach was mimicked in the study to even improve the classification accuracy obtained from the last analyses, as shown in previous section. It led to a one-point increase in global accuracy.

2.7. Global scheme of the health estimation process

The global health estimation process, presented in Figure 4, is semi-automatic as the expert remains present in the process, to analyze and recommend maintenance actions for transformers whose Health Index is evaluated at 3, the most critical level. It is also hybrid because it relies on a machine learning core and on rules provided by the experts in the field.

It consists of three blocks:

- The machine learning prediction, based on the last analysis.
- The legal rules that ensure the compliance of any parameter of this last analysis.
- The expert rules that consider the evolution in dissolved gas concentration evolution through time, based on previous analyses.

The details of the process that includes those three main blocks are provided in Figure 4. In this figure, note that details about the cost matrix are provided in section 3.

Finally, once HI has been estimated, if it appears to be equal or superior to 2, an expert is asked to review the analyses and to provide recommendations in terms of maintenance and/or additional analyses to perform. It can also trigger a reclassification to class 1 (healthy transformer) if the expert concludes that a HI of 2 or 3 is not justified (Figure 5).

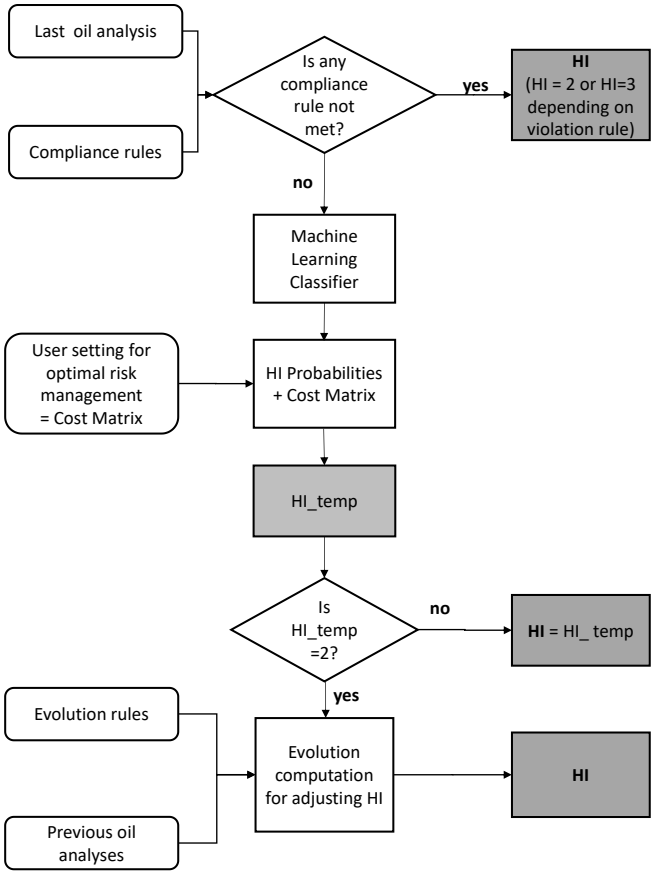


Figure 4: HI estimation synoptic. Inputs are on the left; computation are in the middle, and output is on the right.

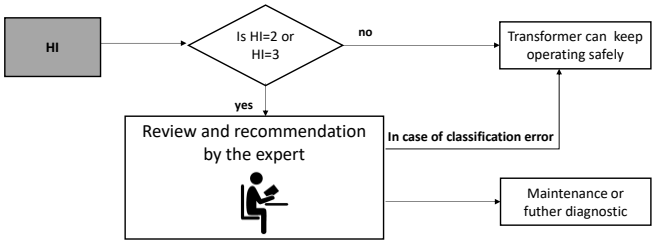


Figure 5: Health Index usage with "expert in the loop"

3. RISK MANAGEMENT FEATURE IN HEALTH ESTIMATION

3.1. Methodology

As described in the introduction, not detecting an at-risk transformer would lead to a risky situation that could have catastrophic consequences. On the other hand, wrongly categorizing a transformer as at-risk whereas it is healthy has minor consequences, as any at-risk transformer will be manually expertised. Indeed, the expert who would analyze such a transformer would set it back in the correct healthy category, with minor consequences, except the time spent for such analysis and action. This means that there is a different cost associated to false positive (healthy transformers wrongly detected as at-risk ones) and false negative (at risk transformers wrongly detected as healthy ones), with false negative being more penalizing than false positive. Also, such asymmetry depends on the context of usage. Indeed, having an at-risk transformer into the wild is always a situation to be avoided, but in some cases, it could be even more damaging than in other cases, considering the criticality of the systems that are supplied by the transformer (hospitals for instance) and considering the environment of the transformer and the risks in case of fire. In this context the goal is now to optimize the classification algorithm not regarding global accuracy, but to a cost function that considers various levels of false positive and false negative costs. Practically, we proceeded with the following steps:

- A cost matrix is defined, attributing some arbitrary weight to each of the errors, reflecting the higher cost of errors for false positive vs. false negative.
- From the classifier, the likelihood that the HI is 1, 2, or 3 is extracted thanks to the Random Forest that easily outputs a calibrated likelihood for each class.
- Considering the likelihood for each class and the cost of each possible choice for prediction, the HI is chosen so that it maximizes the total cost.

Figure 6 shows an example with three different settings (low, medium and high costs), with higher and higher cost for false negative. The goal of higher cost is to better detect at-risk transformers. Technically speaking, the goal is to increase the recall of at-risk transformers (HI=2 and, even more, HI=3). In this example, the likelihood of prediction of each HI is provided. Using a neutral cost matrix, the more likely HI would be selected. In this case HI is predicted as a 1. Using a medium cost matrix, because of the costs, HI is predicted as a 2 as it maximizes the global cost. Similarly, using a stronger cost matrix, named high, HI=3 is selected. The approach was generalized to 15 settings with increasing weights and tested on a 12,000 analyses validation dataset. This led to the results presented in Figure 7 and Figure 8, showing the expected effect on precision and recall: recall for HI=3 class increases as the setting gets more conservative (higher costs) and parallelly its precision decreases. Regarding HI=1 class, its precision increases and its recall decreases as the setting gets more conservative. As an intermediate class, HI=2 sees its precision decrease as the setting gets more conservative. Its recall first increases (as more analyses are correctly classified in HI=2 class, instead of HI=1 class) and then decreases: this because when the setting gets highly conservative the classification tends to incorrectly class HI=2 in the HI=3 class, as it can be seen in the confusion matrix (Figure 9).

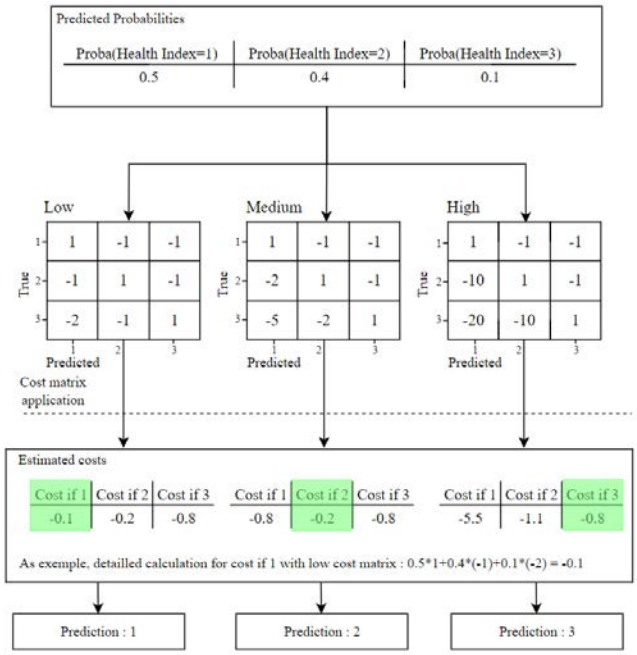


Figure 6: Illustration of cost matrix impact on prediction.

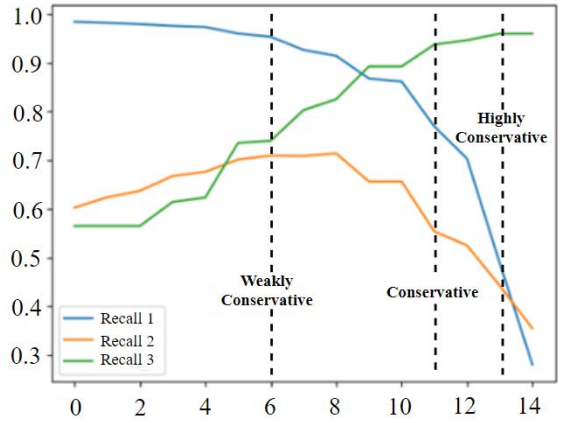


Figure 7: Recall obtained for HI equal to 1, 2 and with using increasingly settings, from 0 to 14.

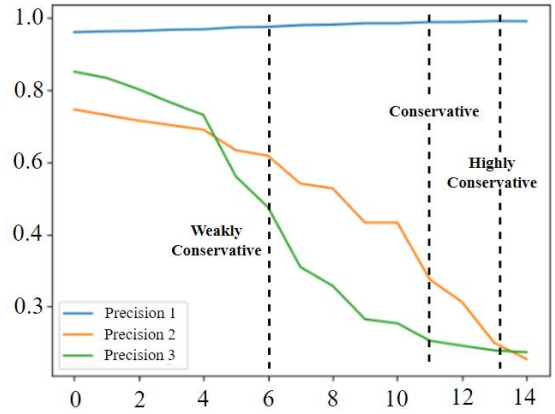


Figure 8: Precision obtained for HI equal to 1, 2 and 3 with using increasingly conservative settings, from 0 to 14.

In order to choose the best setting from the experts' point of view, we randomly selected 200 samples and provided the classification results for all the 15 settings, highlighting the correct results and the wrong ones (correct result means predicted HI equal to the HI estimated by the subject matter expert). This way, subject matter experts could easily see the conservatism level of each setting and choose the three settings that were the most pertinent to them, for covering the global range of criticality of the transformers' context. Those three settings, named weakly conservative, conservative and highly conservative cost matrices, are shown in Figure 7 and Figure 8.

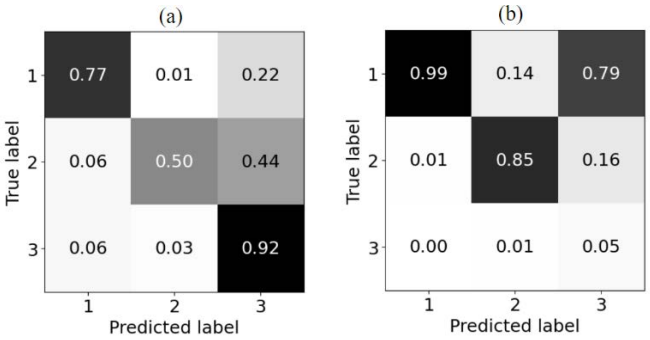


Figure 9: Confusion matrix normalized by rows and by columns for highly conservative cost matrix (respectively (a) and (b)). Results obtained on a 6k analyses dataset.

3.2. Results validation

The global algorithm (described in Figure 4) was tested on a new dataset of 6,000 analyses (called test dataset), using the three settings chosen by the subject matter experts. The results are shown in Table 3. They are consistent with the results obtained on the validation dataset. Particularly, precision for HI=1 class and recall for HI=3 class are very close. For instance, using the conservative cost matrix, precision for HI=1 class is 99% on the test dataset and 99% on the validation dataset, recall for HI=3 class is 95% on the validation dataset and 94% on the test dataset. Those performances and the risk management settings allow for industrial usage.

Table 3: Results obtained on the test dataset for weakly conservative, conservative, and highly conservative settings, focusing on precision for HI=1and recall for HI=3.

Total Number of Analyses=6000	Accuracy	Precision(HI=1)	Recall(HI=3)
Weakly Conservative Cost Matrix	**95%**	**0.98**	**0.75**
Conservative Cost Matrix	**91%**	**0.99**	**0.94**
Highly Conservative Cost Matrix	**76%**	**0.99**	**0.98**

4. CONCLUSION AND DISCUSSION

A semi-automatic, hybrid machine learning and expert-based approach for transformer maintenance has been developed. This approach is based on a very large number of analyses (65,600) carried out over more than 10 years and the technical experts who carried them out. It is called semi-automatic because the expert remains present in the process, especially to analyze and recommend maintenance actions for transformers whose Health Index is evaluated at 3, the most critical level. It is called hybrid because it relies on a machine learning core and on rules provided by the domain experts. The machine learning part goes beyond the application of combinatorial rules: it has captured the experience and practices of experts exposed to many analyses and their practical experience on many transformers, throughout their lifecycle, who are familiar with the signatures of faults and their probability of leading to more serious problems later. Like any machine learning algorithm, the performance of this solution relies on a big amount of data for training. Knowing that this solution is a hybrid solution that also relies on expertise, any industry with advanced expertise necessarily also possesses a large volume of data. Therefore, it is suitable to any industrial player in the field. As in most classification problems, it is not possible to simultaneously improve precision and recall, or in other words, minimize false alarms and minimize non-detections. This is why we have introduced a setting that allows us to prioritize one or the other, depending on the context of use.

Our overall health estimation system relies on:

- This machine learning-based estimation core.
- Legal and safety rules that need to be verified.

Calculations commonly used by experts based on the evolution of dissolved gas concentrations trough time to discriminate the most ambiguous cases.

- The expert who will confirm the critical cases (2 or 3) and provide an appropriate maintenance or further analyses recommendations.

This transformers health estimation enabling predictive maintenance is now deployed on the cloud as an API that is exposed to users whose use can also be done directly through a web application. Today, this process relies on discrete oil analyses; tomorrow, with more and more embedded monitoring in transformers, it is possible to perform real-time analyses. The hybrid approach can be preserved, but this time the machine learning core can rely on the time series and be even more sensitive to any degradation and more accurate in failure prediction.

Finally, it should be noted that this methodology is transferable to many other application domains beyond transformers.

REFERENCES

Quynh T. Tran, Kevin Davies, Leon Roose, Puthawat Wiriyakitikun, Jaktupong Janjampop, Eleonora Riva Sanseverino and Gaetano Zizzo (2020). A Review of Health Assessment Techniques for Distribution Transformers in Smart Distribution Grids, *Applied sciences, 2020, Vol.10 (22), p.8115*

Ahmed E. B. Abu-Elanien, M.M.A. Salama, M. Ibrahim (2012). Calculation of a Health Index for Oil-Immersed Transformers Rated Under 69 kV Using Fuzzy Logic. *IEEE transactions on power delivery, 2012, Vol.27 (4), p.2029-2036*

Khalil Ibrahim, R.M. Sharkawy, H.K. Temraz and M.M.A. Salama (2016). Selection criteria for oil transformer measurements to calculate the Health Index. *IEEE transactions on dielectrics and electrical insulation, 2016, Vol.23 (6), p.3397-340*

P. Sarajcev, D. Jakus, J. Vasilj, M. Nikolic (2018). Analysis of Transformer Health Index Using Bayesian Statistical Models. *2018 International Conference on Smart and Sustainable Technologies, 2018, p.1-7*

Atefeh Dehghani Ashkezari, Hui Ma, Chandima Ekanayake, Tapan K. Saha (2012). Multivariate analysis for correlations among different transformer oil parameters to determine transformer health index. *2012 IEEE Power and Energy Society General Meeting, 2012, p.1-7*

Jahanzaib Javid, Muhammad Ali Mughal, Mustansir Karim (2021). Using kNN Algorithm for classification of Distribution transformers Health index. *2021 International Conference on Innovative Computing (ICIC), 2021, p.1-6*

Ahmed E. B. Abu-Elanien, M. M. A. Salama, Malak Ibrahim (2011). Determination of transformer health condition using artificial neural networks, *2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, p.1-5*

BIOGRAPHIES

Augustin Cathignol is with Schneider Electric. As both a principal Data Scientist and expert in Reliability, monitoring, failure prediction, he is the domain leader for Prognostics and Artificial Intelligence solutions developed by the Artificial Intelligence Hub, for Schneider electric assets and systems. Working closely with subject matter experts, he believes in hybrid solutions that make the most of AI and physics-based models provided by the experts. Prior to joining Schneider Electric, he was for 9 years with Thales-Safran in the field of infra-red detectors development and manufacturing, as the technology reliability manager, and head of the Digital Factory. He spent also 5 years with IBM as a research engineer in the field of reliability of advanced Silicon Devices. He holds a master

degree in Electonics and Telecommunications from *Ecole Supérieure de Chimie Physique Electronique de Lyon*, a Ph.D. in Microelectronics from the *University of Grenoble*, and a master degree in Data Science from *Ecole Polytechnique*, France. Augustin is an expert for the European Commission and member of the French standardization group for Artificial Intelligence.

Hybrid Prognostics for Aircraft Fuel System: An Approach to Forecasting the Future

Shuai Fu*, Nicolas P. Avdelidis

IVHM Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, Bedford, MK43 0AL, United Kingdom

*felix.fu@cranfield.ac.uk
np.avdel@cranfield.ac.uk

ABSTRACT

The copious volumes of data collected incessantly from diverse systems present challenges in interpreting the data to predict system failures. The majority of large organizations employ highly trained experts who specialize in using advanced maintenance equipment and have specific certification in predictive maintenance (PdM). Prognostics and health management (PHM) connects research on deterioration models to strategies for PdM. Prognostics refer to the process of estimating the time until failure and the associated risk for one or more current and potential failure modes. Prognostics aim to provide guidance by alerting to imminent failures and predicting the remaining useful life (RUL). This eventually leads to improved availability, dependability, and safety, while also reducing maintenance costs. This research work diverges from existing established prognostic methodologies by emphasising the use of hybrid prognostics to predict the future performance of an aircraft system, especially the point in which the system will cease to operate as intended, often referred to as its time to failure. We have developed a new method that combines a physics-based model with the physics of failure (PoF) and a multiple-layered hyper-tangent-infused data-driven approach. The results are useful. The authors retrieved datasets for analysis using a laboratory aircraft fuel system and simulation model. Consequently, the comparative results demonstrate that the proposed hybrid prognostic approach properly estimates the RUL and demonstrates strong application, availability, and precision.

Keywords: health management; physics of failure; hybrid prognostics; aircraft fuel system; remaining useful life.

1. INTRODUCTION

The goal of prognostics is to accurately detect and report impending system failures—that is, to forecast the progression of failure. Prognostic methodologies used in prognostic and health management (PHM) achieve this objective through three distinct classifications: condition-based, usage-based, and traditional. Traditional prognostic approaches can be further classified as model-based, data-driven, or hybrid models (Gu & Pecht, 2008; Liao & Köttig, 2014).

Using failure physics (PoF), likelihood, and reliability models to come up with and use expressions is what model-based prognostic methods do. These models utilise the relationships between materials, manufacturing processes, and the dependability, robustness, and strength of a subsystem. This is typically achieved through controlled, structured experiments and life evaluations. Although modelling offers the potential for high accuracy, its implementation and utilisation in complex operational systems are difficult. The models comprise acceleration factor-incorporated reliability testing models, probability models, distributions, and reliability theory principles. Figure 1 shows the comparison between physics-based and traditional condition-based data (CBD) approaches to PHM.

Data-driven prognostic approaches, such as statistical and machine learning methods, are easier to use than model-based approaches but may result in less precise and accurate prognostic projections (Galar et al., 2021). As shown in Fu et al. (2023), statistical approaches include both parametric and nonparametric models. They also include K-nearest neighbour (KNN), a nonparametric method for classifying or regressing an item based on its nearby data points. Linear discriminant analysis (LDA) sorts many objects into groups, hidden Markov modelling (HMM) deals with systems that have hidden states, and principal component analysis (PCA) changes variables in a straight line. Hybrid approaches combine model-based and data-driven methods to enhance

accuracy and gain a deeper understanding of the interactions between parameters and objects. The complexity of computational processing is one of the limitations.

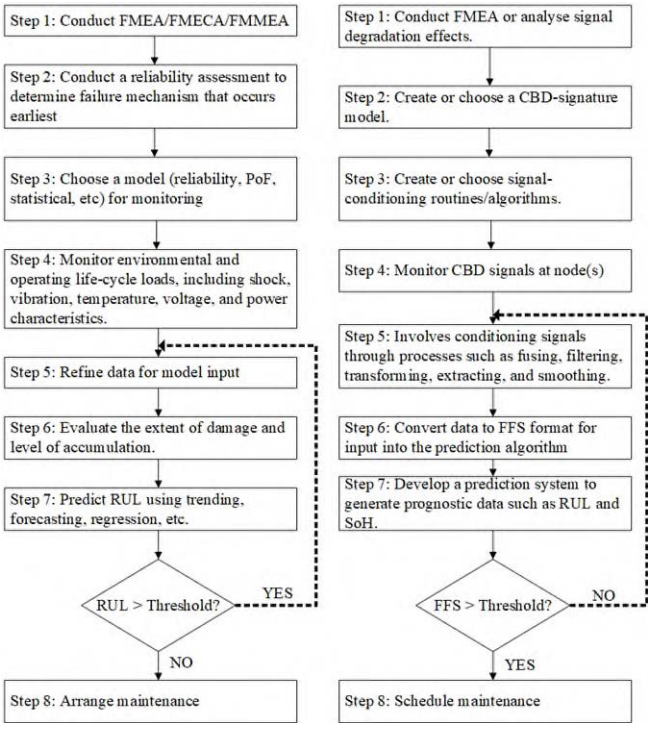


Figure 1. Diagram comparison of model-based and CBD-signature approaches to PHM (Hofmeister et al., 2017).

Figure 2 illustrates an alternative representation of a fault tree for aircraft fuel error-identified systems. Failure Mode and Effects Analysis (FMEA) and Failure Mode, Effects, and Criticality Analysis (FMECA) are used in this study to investigate a fuel-error defect and find the most likely failure mode. This could be air flow, pressure, temperature, or the fuel pump.

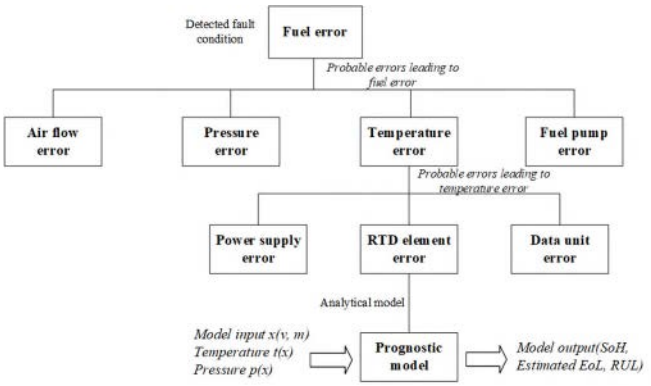


Figure 2. Example of fuel error leading to the application of prognostic model (Douglas Goodman et al., 2019)

In this instance, we conclude that a temperature inaccuracy is the probable factor responsible for the fuel error. The fault tree indicates that three failure modes, namely the power supply, resistive temperature detector (RTD), or an air-data unit, are likely to cause the temperature error. We use an appropriate analytical model to generate prognostic data in the event of an RTD failure.

1.1. Hybrid prognostic mechanisms

A hybrid technique combines physics-based and data-driven prognostics in two phases: offline and online. The initial phase involves creating the nominal and deterioration models, as well as establishing the faults and performance criteria required to predict the remaining useful life (RUL) of the system. The second phase entails using models and thresholds to identify fault initiation, assess the state of system health (SoH), and forecast future SoH and RUL. Data from experiments and synthetic datasets from simulations that replicate real-world settings often validate and optimise the models. We create and utilise sensors to gather data from operational systems, with the aim of monitoring and maintaining the systems' health. The hybrid model offers a higher level of precision compared to employing solely a physics-based or data-driven approach. A physics-based model generates particularly accurate prognostic information when adjusted to sensor data. One drawback is the increased complexity involved in adapting the model to sensor data.

Hybrid models utilise a blend of multiple models to enhance accuracy. Many academics have overlooked hybrid modelling for fault diagnostics and maintenance decision-making. Ahmadzadeh & Lundberg (2014) examined three advanced models for predicting RUL: knowledge-based models, data-driven models, physics-based models, and hybrid prognostic models. Jardine et al. (2006) conducted an examination of machinery diagnostics and prognostics, showcasing the application of statistical, artificial intelligence, and physics-based prognostic methods in condition-based maintenance (CBM) to improve the precision of equipment RUL estimation. A few studies have especially concentrated on hybrid prognostic approaches to capitalise on the benefits of several prognostic models.

Hybrid prognostic methodologies have limitations because they rely on both model-based and data-driven methods. Inaccurate models, noisy data, or both may result in an incorrect RUL forecast. As a result, if not managed correctly, there is a significant likelihood of increased variance in mistakes. A hybrid strategy combines elements of physics-based and data-driven methodologies to leverage their advantages while mitigating their limitations, but it still retains some disadvantages of both. Elattar et al. (2016) developed a flowchart to assist in choosing a prognostic method, as shown in Figure 3.

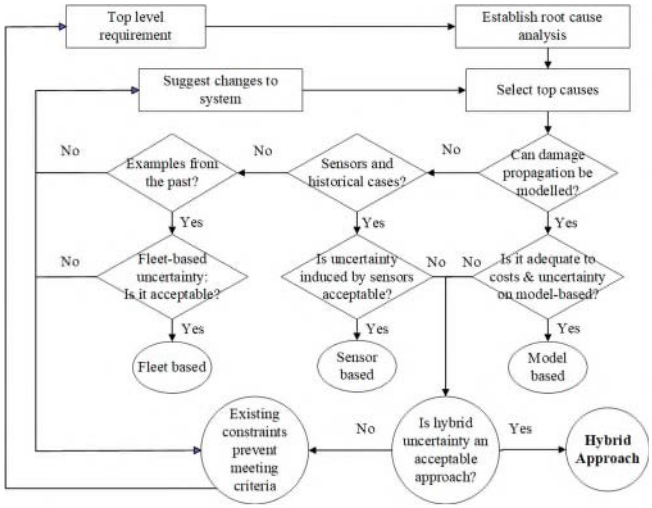


Figure 3. Workflow to select prognostic approaches (Elattar et al., 2016).

A hybrid strategy can effectively integrate data-driven and physics-based methods to optimise their respective strengths when managed appropriately. A physics-based method can address data deficiencies, while a data-driven model can address gaps in understanding the system's mechanics. Performing this fusion before estimating the RUL is known as pre-estimation. Fusion is a process that combines the results of various methods to determine the final RUL after predicting it. Li et al. (2019); Nieto et al. (2016); and Orsagh et al. (2003) used a fusion strategy for aircraft engine bearings to show that this method gives more accurate and long-lasting results than just using data-driven or physics-based approaches alone.

1.2. Prognostic application

In the aircraft sector, there are several instances of prognostic applications that are now in the developmental stage. The current aim of prognostic society is to create a PHM system capable of detecting and isolating problems in both the primary and subsystems of the aircraft. Additionally, this system will offer prognostic information for specific components (Losik, 2012; McCollom & Brown, 2011; Vohnout et al., 2012). PHM, which is critical to improving safety and lowering maintenance costs, has a significant impact on the choice of aircraft. The proposed architecture incorporates an external PHM system that will employ data mining techniques. Figure 4 depicts the forecasting applications.

There has been a notable surge in interest in prognostics due to their ability to improve the health management of intricate engineering systems. Prognostics are important because they allow us to predict future illness progression and treatment outcomes. Daily weather forecasting also employs this technology. Whether they are located on board or off board, prognostic software solutions have the potential to function in real-time or nearly real-time.

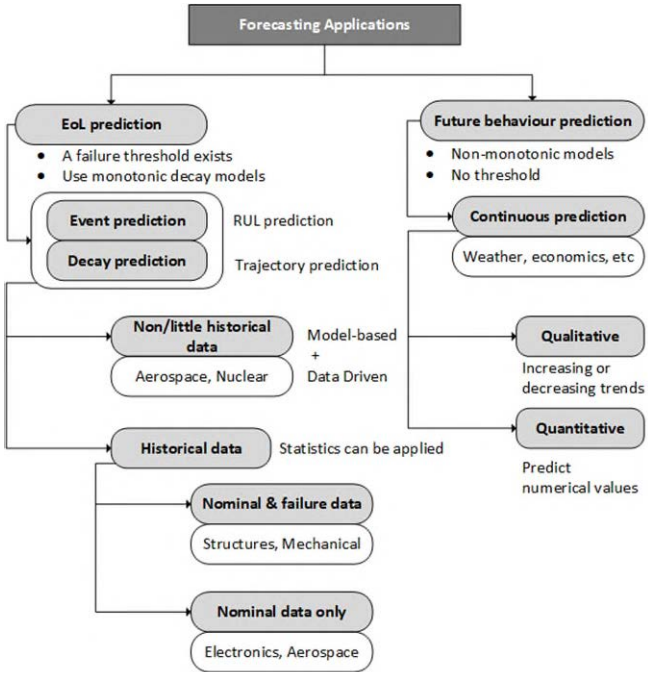


Figure 4. Forecasting applications.

Prognostics can be used offline, regardless of how long the monitored system has been in operation. Real-time prognostics uses the online data collected from the data collection system to accurately estimate the RUL and warn about an imminent breakdown. This allows the system to be reconfigured and the mission re-planned. The offline prognostics system utilises extensive system data from the whole fleet and applies intricate data analysis techniques that are not feasible to conduct in real-time on board due to resource and time constraints. An offline prognostic system in logistical support management can provide useful information for maintenance planning and decision-making.

2. AIRCRAFT FUEL DELIVERY SYSTEM

An aircraft fuel delivery system with three tanks usually consists of a central tank and two wing tanks. The central tank supplies fuel to the engine, and the wing tanks supply fuel to the central tank via pumping stations. Two centrifugal pumps, complete with check valves to prevent backflow, equip each station. Prime movers, operating at a constant angular velocity, power these pumps. Engineers designed the system with varying elevations between the tanks and the engine intake to facilitate fuel flow. This flow is regulated by two-way bidirectional valves that respond to the fuel levels in each tank. Figure 5 illustrates the design of the simulation model, which is based on the MathWorks library.

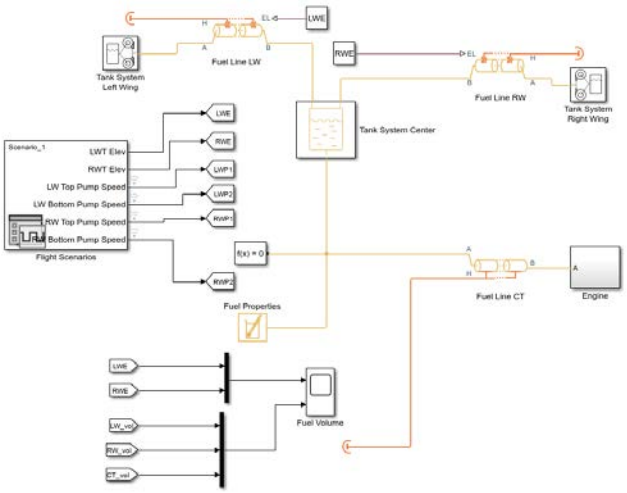


Figure 5. Aircraft fuel delivery simulation system.

The process of simulating this system involves modulating the fluid dynamics associated with the fuel flow, the mechanical design of the pumps and valves, and the control systems responsible for overseeing fuel distribution. The operation will examine the effects of aircraft manoeuvres, specifically changes in bank angle, on the reduction in pressure across the fuel lines. Figure 6 depicts the structure of the central tank in the simulation model.

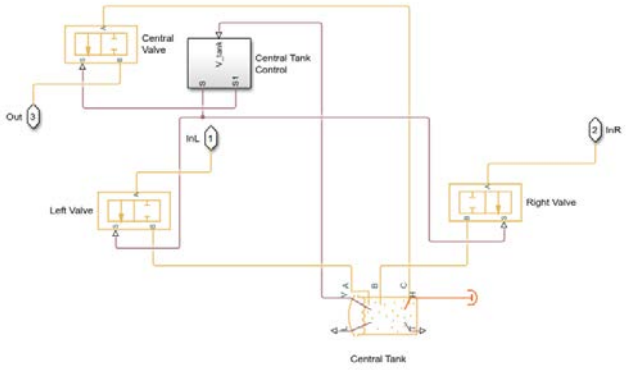


Figure 6. Central tank structure.

There is a storage tank in a thermal liquid network that maintains a constant pressure and allows for a variable number of inlets. The pressure at the liquid surface is considered to be equivalent to the pressurisation. It represents the hydrostatic pressure differential between the fuel surface and the inlets. When the liquid level drops below the inlet height, the port is exposed. It is connected to a partially filled pipe to simulate the ongoing decrease in liquid level within the pipe. In the simulation model, ports A, B, C, D, E, and F are thermal liquid conservation ports connected to the tank inlets. The thermal-conserving port H is associated with the liquid's temperature in the tank. The physical signals V, L, and T represent the liquid volume, liquid level, and liquid temperature, respectively. Bidirectional valves are also depicted within a thermal fuel network. The voltage input S determines the location of the spool. Positive spool displacement facilitates fuel flow by opening the connection between ports A and B. The disconnection is caused by reverse spool movement. We regard the aforementioned component as adiabatic. The system does not transfer thermal energy to its surroundings. Figure 7 illustrates the engine pump and its various subcomponents.

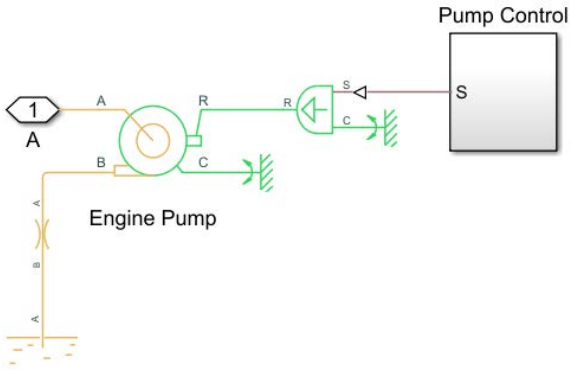


Figure 7. Engine pump and its subcomponents.

The simulation model also includes a centrifugal pump operating within the fuel supply system. We employ affinity laws to establish the relationship between the reference pump characteristics and the actual flow rate and pressure gain. We connect the thermal fuel conservation ports, identified as ports A and B, to the pump's input and outflow, respectively. The drive shaft and casing respectively connect to the mechanical rotational conserving ports, denoted as ports R and C. Mechanical orientation determines the shaft rotation for proper pump functioning, where the flow moves from port A to port B and the pressure increases. The pump's performance in the other direction is indeterminate and perhaps imprecise. Figure 8 shows the engine pump's various characteristics.

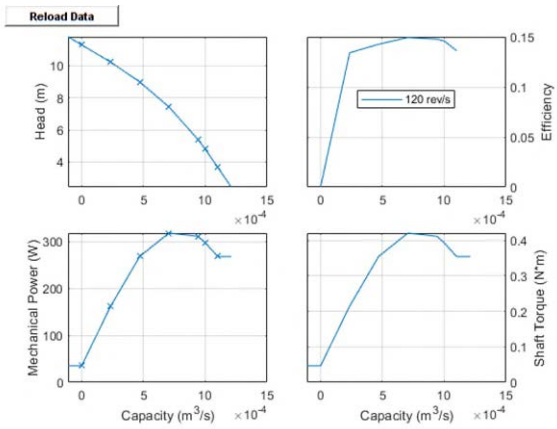


Figure 8. Engine pump characteristics.

It is also possible to get an ideal source of angular velocity from the system, which produces a velocity difference at its

ends that is proportional to the physical input signal. The source is considered ideal since it is thought to have sufficient power to sustain a defined velocity regardless of the torque applied to the system. The relative velocity is calculated by subtracting the absolute angular velocity of the terminal $C(W\_C)$ from the absolute angular velocity of the terminal $R(W\_R)$, denoted as $W = W\_R - W\_C$.

3. HYBRID PROGNOSTIC MODELLING AND RESULTS

3.1. Physical principles

The simulation model suggests a generic simulation of an aircraft's fuel supply system rather than directly replicating a specific real-world aircraft's fuel system. The model encompasses common components of an aviation fuel system, including:

- Multiple fuel tanks: commercial and military aircraft typically have wing tanks and a centre tank to evenly distribute weight and improve fuel economy. The model provides certain starting pressurisation and volume capacities for the tanks, essential for maintaining fuel flow under different flying conditions.
- The specifications for centrifugal pumps and valves, such as bidirectional valves and check valves, demonstrate the intricate systems used to control fuel supply from the tanks to the engines.
- It contains essential information regarding the fuel line's length, diameter, and resistance properties, crucial for accurately modelling fuel flow within the system.

The simulation provides a valuable resource that can be customised or expanded to accurately replicate the fuel system of a particular aircraft. Factors such as tank sizes, pump capacity, and system layout may be able to be adjusted in accordance with the aircraft's technical requirements. Table 1 presents the initial circumstances and parameters of the model.

The physics-based model often involves the monitoring of many parameters, including pressures, temperatures, fuel levels, flow rates, and valve functioning. These parameters are determined by the components involved, as well as their established failure modes. While conducting an analysis of a simulated aircraft fuel system, researchers strive to identify consistent patterns in:

- Fuel consumption rates during comparable operating conditions. Substantial variances could indicate inefficiencies or deterioration, such as pressure or temperature fluctuations in tanks or fuel lines that differ from the usual values, signalling possible problems.
- Valves and pumps have operational behaviour, including unforeseen operations or alterations in performance measurements.

Table 1. The initial circumstances and parameters of the simulated aircraft fuel delivery model

Components	Parameters	Specs
Initial Conditions	Temperature	333.15 K
	Pressure	0.1 MPa
Fuel Tanks	Pressurisation	0.1 MPa
	Minimum fuel volume	0.09463525 m^3
Wing Tanks	Initial volume	10 m^3
	Maximum capacity	12 m^3
Centre Tank	Pressurisation	0.1 MPa
	Initial volume	5 m^3
	Maximum capacity	284 m^3
Pumps	Reference density	920.027 kg/m^3
	Reference angular velocity	120 rev/s
	Angular velocity threshold	10 rad/s
	Operational ranges for angular velocity	0 to 200 rev/s
	Mover time constant	0.2 s
Valves	Maximum opening area	$\pi/4\times(0.03048)^2$ m^2
	Leakage area	1e – 10 m^2
	Cutoff time constant	0.1 s
	Maximum valve opening (2-Way directional valves)	5.1e-3 m
Fuel line piping	Length	5m
	Hydraulic diameter	3.05e-2 m
	Aggregate equivalent length for local resistances	2.56 m

To develop a physics-based model for an aviation fuel system, one needs to understand the basic concepts of fluid dynamics and the mechanical operations of these systems. Common mathematical formulas and concepts are summarised to reflect the physics of aviation fuel systems, establishing a solid foundation for developing a physics-based model.

The principle of mass conservation is applicable to the process of fuel transfer between tanks and its subsequent use by an airplane's engines. The generic equation provided can be used to analyse each tank:

$$\frac{dV}{dt} = Q_{in} - Q_{out} \quad (1)$$

Where:

- V is the volume of fuel in the tank.
- t is time.
- Q_{in} is the inflow of fuel into the tank, and
- Q_{out} is the outflow rate of fuel from the tank to the engines or to other tanks.

The application of Bernoulli's equation, which establishes a relationship between the pressure, velocity, and height head of the fluid, can aid in the analysis of fluid flow between tanks. This is especially beneficial when the tanks are located at varying heights or when calculating the necessary pressure for fuel transfer between them.

$$P + \frac{1}{2}\rho\nu^2 + \rho gh = constant \quad (2)$$

Where:

- P is the pressure within the fluids.
- ρ is the density of the fluid (fuel).
- ν is the velocity of the fluid.
- g is the acceleration due to gravity, and
- h is the height of the fluid column (which could represent the fuel level in the tank).

PID (proportional-integral-derivative) control effectively manages pump speeds or valve positions to maintain predetermined fuel levels in individual tanks. Fuel level sensors provide the input for this control mechanism. The conventional arrangement of a PID controller is as follows:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau + Kd\frac{de(t)}{dt} \quad (3)$$

Where $u(t)$ denotes the control signal (e.g., pump speed), $e(t)$ denotes the error signal (difference between desired and actual fuel level), and K_p, K_i, and K_d denote the proportional, integral, and derivative gains, respectively.

In aircraft fuel systems, where turbulent flow is common, we can use the Darcy-Weisbach equation to calculate the pressure drops (ΔP) along a pipe length:

$$\Delta P = f\frac{L}{D}\frac{\rho v^2}{2} \quad (4)$$

The variables are defined as follows: f is the friction factor, L is the length of the pipe, D is the diameter of the pipe, ρ is the density of the fuel, and v is the velocity of the fuel.

3.2. Hybrid prognostic integration methodology

According to Chao et al. (2021), provided $X_{s_i} = \left[x_{s_i}^{(1)}, \ldots, x_{s_i}^{(m_i)}\right]^T$ are multivariate time-series data from condition monitoring sensors and their accompanying RUL $Y_i = \left[y_i^1, \ldots, y_i^{m_i}\right]^T$ for a fleet of N units $(i = 1, \ldots, N)$. Each observation $x_{s_i}^{(t)} \in R^p$ consists of a vector of ρ raw measurements taken at operating conditions $\omega_i^{(t)} \in R^s$. The length of the sensory signal for the i^{th} unit is determined by m_i, and may vary between units.

The overall cumulative length of the available data collection is $m = \sum_{i=1}^{N} m_i$. We designate the provided dataset more compactly as $D = \left\{W_i, X_{s_i}, Y_i\right\}_{i=1}^{N}$. The objective is to develop a predictive model $\mathcal{G}$ that can accurately estimate the RUL $(\widehat{Y})$ on a test dataset $D_{T*} = \{X_{sj*}\}_{j=1}^{M}$ consisting of M units, which $X_{sj*} = \left[x_{sj*}^1, \ldots, x_{sj*}^{k_j}\right]$ are multivariate time series of sensor measurements. The overall cumulative length of the test data set is $m_* = \sum_{j=1}^{M} k_j$.

The subsequent subsections provide a comprehensive analysis of each of these phases. Eker et al. (2019) proposed the following input and output processes for physical-based approaches, as depicted in Figure 9.

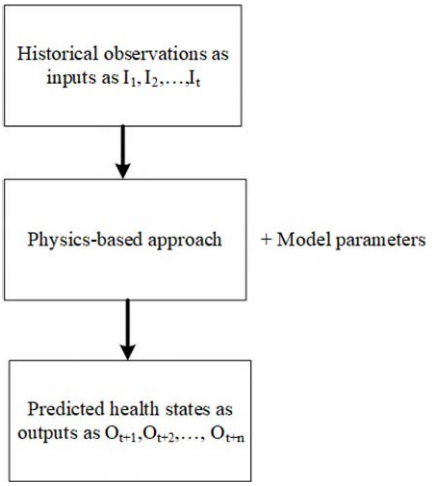


Figure 9. Input and output of the physics-based model.

The flowchart depicted in Figure 10 illustrates the operational mechanism of a hybrid predictive strategy employed in aeroplane fuel distribution systems. This approach combines physics-based and data-driven models. The commencement of the process occurs subsequent to the identification of the components and modes of aircraft failure, the process commences. The aviation fuel system is analysed using physics-based techniques and domain expert knowledge to estimate the short-term RUL. We conduct the analysis using either a real-world or synthetic dataset.

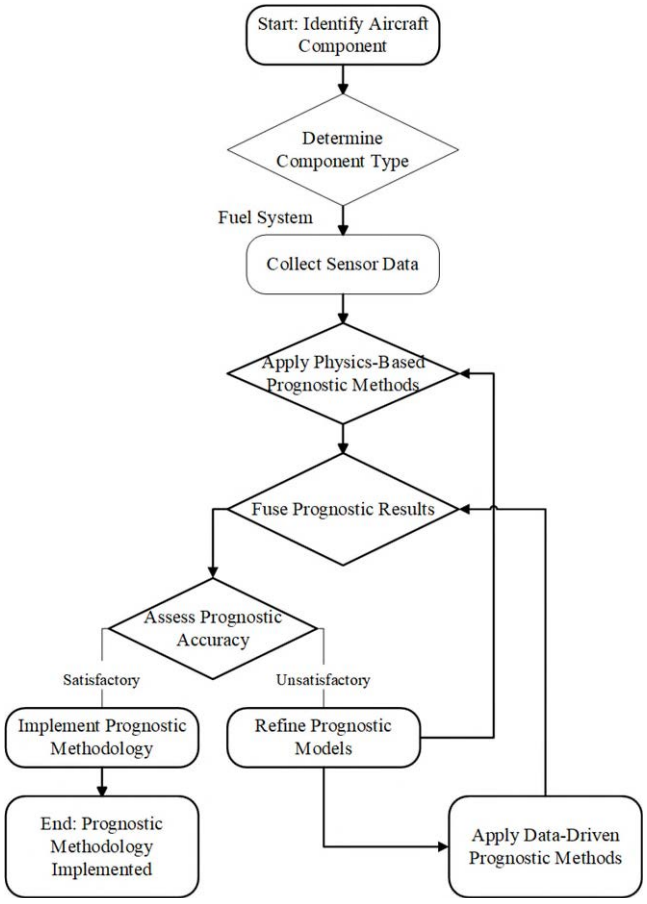


Figure 10. Illustration of fusion mechanism of a hybrid prognostic methodology for an aircraft fuel system

The prediction results will be assessed using several prognostic metrics, as demonstrated in the studies conducted by Chao et al. (2021) and Fu & Avdelidis (2023). The optimisation process will be carried out by comparing the accuracy results with the actual RUL. Once the desired outcome is attained, a hybrid prognostic technique will be included. As long as the engineering systems adhere to specific physical deterioration, the methodology flowchart can be used for other complicated systems.

Random holdback is the chosen approach for validation. The neural network consists of two layers in total. In the initial layer, three radial Gaussian activations are employed, whereas the subsequent layer utilises two times Sigmoid TanH and a linear activation function, which bears a striking resemblance to the activations used in two-layer models. We set the learning rate at 0.1, allowing for robust fitting. A single round of a tour is subject to a penalty approach. The authors of this paper used a variety of neural networks with different activation functions, including Sigmoid TanH, identity linear, and radial Gaussian. There are variations in the outcomes observed among the different models. Various characteristics were obtained, and the highest-ranked attributes that have the most impact on achieving the best result were selected. Figure 11 depicts the simplified neural network that yields the best results.

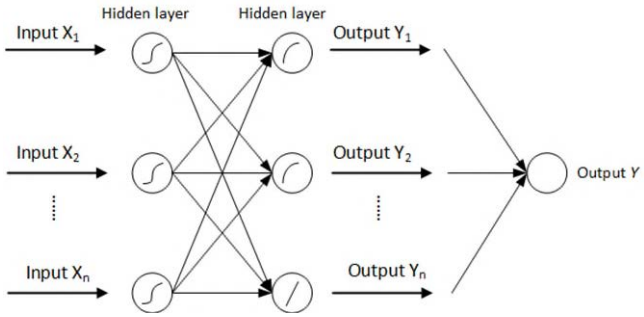


Figure 11. Simplified boosted neural network model NGaussian(3) NTanH(2)NLinear(1)

In its fitting routine, the boosted neural network employs a validation mechanism. Validation methods include holdback, K-fold, or the use of a validation column that performs the following actions to fit the model:

- The model parameters are subject to a penalty.
- The validation set adjusts the penalties applied to the parameters.

The actual-by-predicted plot calculates the comparison between the training's actual and expected values. The suggested methodology reveals the correlation between the observed value and the projected value on the training dataset. The ideal situation involves aligning all data points along a straight path where the anticipated values accurately match the observed values. The data points in this graph display a mostly linear trend that is primarily located close to the actual RUL value. This observation suggests that the projected values exhibit a degree of resemblance to the observed values. As a result, the model exhibits higher levels of predicted accuracy for positive values in comparison to negative values. Table 2 presents the results for both training and validation methods.

Table 2 suggests multiple measuring and evaluation measures to compare prognostic outcomes. Fu et al. (2023) provide comprehensive explanations for each rating metric. Table 2 demonstrates that the R^2 value is 0.9998 for both the training and validation stages, indicating a substantially identical outcome in both phases. The training procedure yielded a higher RASE value of 14.56 compared to the validation process value of 21.26, indicating that the prognostic algorithm exhibits superior performance during the training phase as opposed to the validation phase. We may attribute the tiny difference to the inadequate amount of training data, which led to less accurate predictions. Future optimisation and updates have significant potential to improve accuracy. MathWorks extracts the simulation data from the simulated fuel distribution systems, which you can view at https://zenodo.org/doi/10.5281/zenodo.10888497.

Table 2. Optimal variation in terms of evaluation performance.

	Measures	Value
Training	RSquare	0.9997863103
	RASE	14.563087252
	Mean Abs Dev	8.6751876436
	-LogLikelihood	2308.3148678
	SSE	127038.02268
	Sum Freq	599
Validation	RSquare	0.9997561214
	RASE	21.260680304
	Mean Abs Dev	9.1951405963
	-LogLikelihood	1173.5466987
	SSE	135604.9581
	Sum Freq	300

4. CONCLUSION

Prognostics are essential in PHM, comprising several elements like system monitoring, fault detection and diagnostics, failure prognostics, and operating management. Prognostic models in both industry and research commonly utilise physics-based and data-driven methodologies. Every strategy has unique benefits and drawbacks. The current work presents a hybrid prognostic model that efficiently incorporates the benefits of both approaches while reducing their limits whenever possible.

Hybrid prognostics were modified in order to incorporate the short-term forecast from physics-based prognostics. This concept has been used in aviation fuel distribution systems. The present research compares the RUL estimations achieved by the hybrid method with those acquired through several physics-based and data-driven methodologies. In real-world scenarios with insufficient data on long-term failures, the hybrid strategy significantly outperforms any of its component techniques.

ACKNOWLEDGEMENT

This research has received funding from the European Commission under the Marie Skłodowska Curie program through the H2020 ETN MOIRA project (GA 955681). Authors gratefully acknowledge the research team at IVHM Centre, Cranfield University, UK, for supporting this research.

REFERENCES

Ahmadzadeh, F., & Lundberg, J. (2014). Remaining useful life estimation: review. *International Journal of System Assurance Engineering and Management*, *5*(4), 461–474. https://doi.org/10.1007/s13198-013-0195-0

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 1–14. https://doi.org/10.3390/data6010005

Douglas Goodman, James P. Hofmeister, & Ferenc Szidarovszky. (2019). *Prognostics and health management : a practical approach to improving system reliability using conditioned-based data*.

Eker, O. F., Camci, F., & Jennions, I. K. (2019a). A New Hybrid Prognostic Methodology. In *International Journal of Prognostics and Health Management*.

Eker, O. F., Camci, F., & Jennions, I. K. (2019b). A new hybrid prognostic methodology. *International Journal of Prognostics and Health Management*, *10*.

Elattar, H. M., Elminir, H. K., & Riad, A. M. (2016). Prognostics: a literature review. *Complex & Intelligent Systems*, *2*(2), 125–154. https://doi.org/10.1007/s40747-016-0019-3

Fu, S., & Avdelidis, N. P. (2023). Prognostic and Health Management of Critical Aircraft Systems and Components: An Overview. *Sensors*, *23*(19). https://doi.org/10.3390/s23198124

Fu, S., Avdelidis, N. P., & Jennions, I. K. (2023). A Prognostic Approach to Improve System Reliability for Aircraft System. *2023 7th International Conference on System Reliability and Safety (ICSRS)*, 259–264. https://doi.org/10.1109/ICSRS59833.2023.10381117

Galar, D., Goebel, K., Sandborn, P., & Kumar, U. (2021). *Prognostics and Remaining Useful Life (RUL) Estimation: Predicting with Confidence*. https://doi.org/10.1201/9781003097242

Gu, J., & Pecht, M. (2008). Prognostics and health management using physics-of-failure. *2008 Annual Reliability and Maintainability Symposium*, 481–487. https://doi.org/10.1109/RAMS.2008.4925843

Hofmeister, J., Szidarovszky, F., & Goodman, D. (2017, March). *AN APPROACH TO PROCESSING CONDITION-BASED DATA FOR USE IN PROGNOSTIC ALGORITHMS*.

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. https://doi.org/10.1016/J.YMSSP.2005.09.012

Li, Z., Wu, D., Hu, C., & Terpenny, J. (2019). An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. *Reliability Engineering and System Safety*, *184*, 110–122. https://doi.org/10.1016/j.ress.2017.12.016

Liao, L., & Köttig, F. (2014). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, *63*(1), 191–207. https://doi.org/10.1109/TR.2014.2299152

Losik, L. (2012). Using the prognostic health management program on the air force next generation reusable launch vehicle. *AIAA SPACE Conference and Exposition 2012*.

McCollom, N. N., & Brown, E. R. (2011). PHM on the F-35 fighter. *2011 IEEE International Conference on Prognostics and Health Management, PHM 2011 - Conference Proceedings*. https://doi.org/10.1109/ICPHM.2011.6024363

Nieto, P. J. G., García-Gonzalo, E., Sánchez, A. B., & Fernández, M. M. (2016). A new predictive model based on the abc optimized multivariate adaptive regression splines approach for predicting the remaining useful life in aircraft engines. *Energies*, *9*(6). https://doi.org/10.3390/en9060409

Orsagh, R. F., Sheldon, J., & Klenke, C. J. (2003). Prognostics/diagnostics for gas turbine engine bearings. *IEEE Aerospace Conference Proceedings*, *7*, 1165–1173. https://doi.org/10.1109/AERO.2003.1234152

Vohnout, S., Bodden, D., Kim, B. U., Wagoner, R., Kunst, N., Edwards, P., Gleeson, B., Cascio, D., Brzuszkiewicz, S., Wagemans, R., Rounds, M., & Clements, N. S. (2012). Prognostic-enabling of an electrohydrostatic actuator (EHA) system. *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2012, PHM 2012*, 437–448.

Influence of Reducing the Load Level of Mission Profiles on the Remaining Useful Life of a TO220 Analyzed with a Surrogate Model

Tobias Daniel Horn[1], Jan Albrecht[2] and Sven Rzepka[3]

[1]*Fraunhofer ENAS, Technologie Campus 3, 09126 Chemnitz*
tobias.horn@enas.fraunhofer.de

[2,3]*Fraunhofer ENAS, Technologie Campus 3, 09126 Chemnitz and*
Chemnitz University of Thechnology, Center for Micro and Nano Technologies, Reichenhainer Str. 70, 09126 Chemnitz
jan.albrecht@enas.fraunhofer.de
sven.rzepka@enas.fraunhofer.de

ABSTRACT

A methodology for replacing finite element simulations with a fast-calculating surrogate model for fault tolerance in operating systems is presented. The study focuses on the TO220 rectifier system and explores methods to detect impending failures and calculate the resulting necessary load reduction. The finite element simulation model is described, highlighting the die attach as the relevant connection for failure. A surrogate model is developed using long-short-term-memory models to predict temperature and in-elastic strain. The surrogate model significantly reduces simulation time, allowing for the adjustment of load based on the system's current state of health. The rainflow counting algorithm is applied to calculate the number of cycles to failure, and the Palmgren-Miner linear damage accumulation relation is used to determine the damage and state-of-health. The dependency of the change in lifetime due to variations in scaling factor is evaluated and the results show that load reduction increases the lifetime of the system.

1. INTRODUCTION

The increased requirements for fault tolerance (e. g. for SAE level L3 and onward, defined by the Society of Automotive Engineers (SAE) in SAE International (2021)) requires, the operating system must continue to operate with reduced power until other measures are initiated. Therefore, the system must be able to detect the impending failure and start the fault handling. Furthermore, the result of the intervention must be predicted in order to apply right failure rectification. These requirements can be met by various methods, as mentioned by Moeller, Inamdar, van Driel, Bredberg, Hille, Knoll and Vandevelde (2024). For example, a system that regularly undergoes rest phases can run self-diagnoses processes by using standard load cycles during these rest phases. From the deviation of the resulting response to the response in the undamaged state, the damage and the resulting necessary reduction in load can be calculated (as shown in e.g. Chacko, Moeller, Kolas, Albrecht, and Rzepka (2024)). However, the disadvantage of this method is that the fault can be detected at the earliest in the first rest phase after the first measurable deviations have occurred. On the other hand, the advantage is that the calculation must not be carried out in the system itself, but can also be performed in the cloud, for example.

Alternatively, a digital twin of the system can be created and this representation can be digitally loaded in parallel with the real system. The digital twin then calculates the damage to the real system based on the real load. In order to achieve this, the digital twin must be capable of mapping the failure mechanism that occurs and calculating the damage from this. In addition, the calculation of the damage due to the load in the digital twin must be performed faster than the load is applied in reality (this depends on the available calculation resources). Only if both of these conditions are met the current state of health of the system can be mapped correctly and an appropriate regulation can be calculated.

In this work, the chosen system is a TO220 rectifier. The TO220 is a Silicon Carbide Schottky diode for ultra-high performance, low loss, high efficiency power conversion applications. For example, it can be used as a switched-mode power supply, AC-DC and DC-DC converter, in battery charging infrastructure, server and telecommunications power supply, uninterruptible power supply and as a photovoltaic inverter (Nexperia 2023). As already shown by Albrecht, Horn, Habenicht and Rzepka (2023), it is possible to generate a validated digital representation of this rectifier in the form of a combined multi-field FE simulation, from which the damage under real loads can be calculated. However, the calculation time of these FE simulations is

much too long. This paper shows how the FE simulation model can be replaced by a fast-calculating surrogate model. After a brief introduction of the TO220 rectifier and the corresponding FE simulation in section 2, the training and setup of the surrogate model and the calculation of the state of health are presented. Subsequently, this surrogate model is used to calculate the change in lifetime due to the reduction of the load.

2. Finite Element Simulation

The structure of the TO220 rectifier can be seen in the FE simulation model in Figure 1 and Figure 2. The die attach is the relevant connection regarding the failure (as shown in Albrecht et al. (2023)), representing the connection between the chip and the lead frame. The current flows from the contact via the bond wire to the die and is then transferred to the lead frame via the die attach. The materials heat up due to the current flow (Joule heating). The temperature is dissipated via the heatsink.

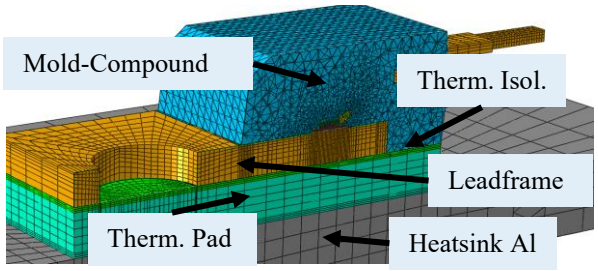


Figure 1: Finite Element model of the TO220 rectifier (cut view).

The FE simulation is based on a sequential approach, where the electric-thermal behavior is simulated first, followed by the thermal-mechanical behavior of the component. A current load profile is used as input for the electric-thermal simulation. The computed temperature field is then used as input for the thermal-mechanical simulation. From the thermal-mechanical simulation the in-elastic strain in the region of the die-attach corners (the relevant area for the failure) is extracted using an averaging approach.

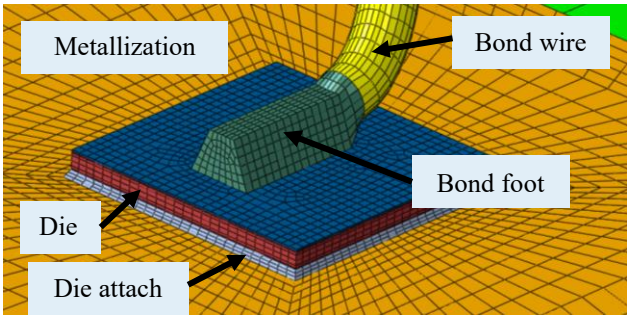


Figure 2: Finite Element model showing the die attach and the bond foot.

In order to calculate the state-of-health from the simulation result, the Coffin-Manson lifetime model

$$N_f = C_1 \Delta\varepsilon_{pl}^{C_2} \quad (1)$$

is used. With this model the number of cycles to failure N_f is calculated by using the in-elastic amplitude allocated to the cycle $\Delta\varepsilon_{pl}$ as well as two model parameters C_1 and C_2. In this calculation, the parameters identified by Darveaux and Banerji (1991) for Pb95Sn5 were used. Since real loads are used in this calculation rather than standard cycles, rainflow counting is carried out based on the temperature profile. The rainflow counting extracts cycles from the real load, and the corresponding change in the in-elastic strain is assigned to these cycles. From the number of cycles to failure N_f the state-of-health can be calculated (as shown in section 4). The methodology of calculating the state-of-health by using FE simulations is also shown in Figure 3.

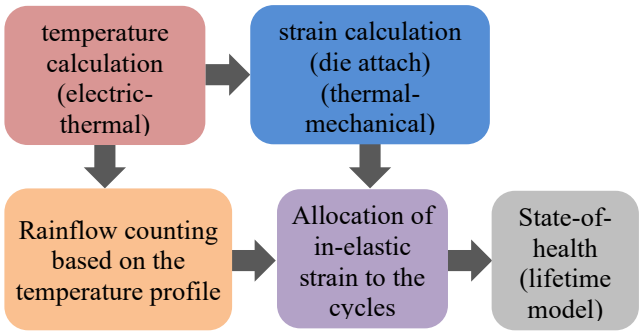


Figure 3: Applied methodology for the calculation of the state-of-health.

A crucial aspect of the FE simulation is the accurate description of the materials used in the component. Most of the materials exhibit strong temperature dependence, the bond wire and bond foot (both aluminum) as well as the lead frame (copper) are modeled by using linear elastic and bilinear kinematic hardening plasticity behavior and the solder in the die attach is modeled by using the Anand law. Especially the solder is highly non-linear and the material behavior depends strongly on the strain experienced in the past. Further details on the materials and the FE simulation in general are shown in Albrecht et al. (2023).

Using the calibrated model, a complete Worldwide harmonized Light vehicles Test Procedure (WLTP) mission profile was simulated by varying the electrical load over time. In order to obtain the current from the WLTP profile, an inverter module was added before the simulation. The temperature field from the electric-thermal simulation was used as input for the thermal-mechanical simulation and the stress as well as the strain in the die-attach corners is calculated and averaged. The results are shown in Figure 4.

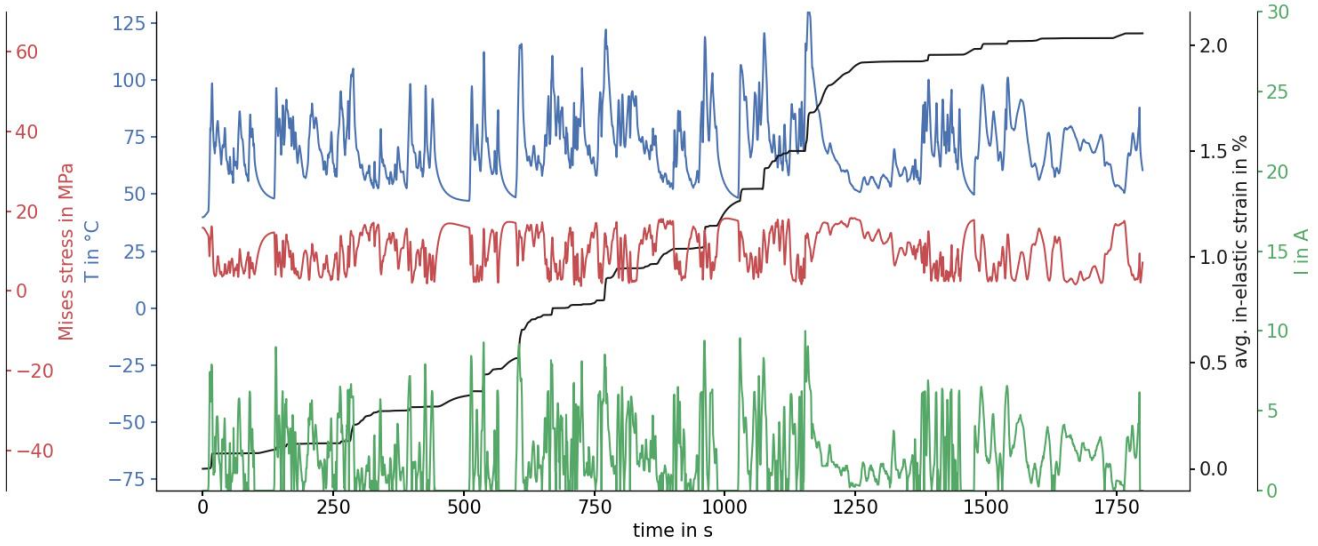


Figure 4: WLTP (current) mission profile (green) as well as some of the calculated results from the FE simulation: temperature (blue), von Mises stress (red) and the averaged total in-elastic strain (black).

3. SURROGATE MODEL

As mentioned before, the WLTP cycle can be simulated by using FE simulation and the results fit to experimental measurements. So, the FE simulation model is a digital representation of the TO220 rectifier. However, the simulation time for calculation the WLTP cycle of 1800s is round about two days. In order to reduce the simulation time, the finite element simulation must be replaced by a surrogate model.

Therefore, the restrictions are: The surrogate model must take the current as input and the temperature as well as the in-elastic strain as outputs. Additionally, the surrogate model must be able to store all information of the reality – and because the FE simulation fits to the reality also all information of the FE simulation model. Due to this, the type of the surrogate model cannot be chosen randomly.

As described before, the materials of the TO220 rectifier are highly non-linear and also strongly dependent to the history. Due to this, as model type the long-short-term-memory (LSTM) model is used (Hochreiter & Schmidhuber (1997)). LSTMs are effective in capturing long-range dependencies in sequential data and have the ability to remember information over long periods of time, as for example shown in Zheng, Ristovski, Farahat and Gupta (2017). Analogous to the FE simulation, two different LSTM models were trained: one model to predict the temperature and one model to predict the in-elastic strain. The training data were produced by the FE simulation model and as the LSTM model is to be applied to real loads, the simulation data from the WLTP cycle is used for training. The training/validation split is 80/20 % and the Adam optimizer (Kingma & Ba 2014) is used. In order to generate additional data that the model had not seen in training, the mission profile was varied using different methods and a total of six variations were calculated using FE simulation. The shown mean absolute error / mean absolute percentage error (MAE/MAPE) is calculated on all seven mission profiles. The models are trained without considering the time. So, for the data a constant time step of 10 Hz is used.

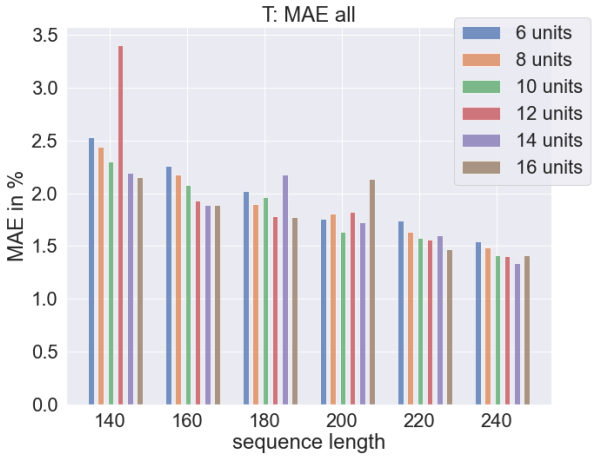


Figure 5: MAPE for variation of the sequence length and the units for the LSTM predicting the temperature.

For the LSTM there are many parameters, such as sequence length, number of unit layers, number of units per layer, features, predictions, learning rate etc. These parameters are optimized by a combination of a variation study and a hyperparameter variation. Exemplarily for the LSTM predicting the temperature, which only uses one unit layer, in Figure 5 the MAPE for the variation of the sequence length and the number of units in the unit layer is shown. The increase of the sequence length (the history taken into

account) significantly increases the prediction quality. Simultaneously, increasing the sequence length reduces the difference between the models with different units.

For predicting the temperature, the LSTM model with the current as feature, a sequence length of 240, one unit layer with 14 units is chosen. The prediction quality (also for all seven mission profiles) is shown in Figure 6.

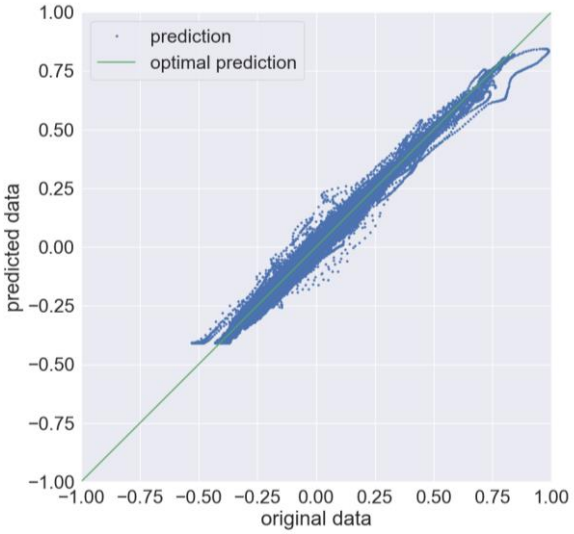


Figure 6: Prediction quality of the LSTM for temperature prediction. The green line indicates the optimal prediction.

As mentioned before, for the prediction of the in-elastic strain a second LSTM model is trained. Therefore, the features are the current together with its first and second derivative and also the temperature (predicted by the other LSTM) together with its first derivative. The result of the parameter variation and hyperparameter variation gives a model with two unit layers with 16 units and 10 units respectively. The sequence length is identified to 200.

The in-elastic strain is a continuously increasing quantity where the changes are constantly added up. Therefore, not the in-elastic strain itself was predicted, but the incremental in-elastic strain. In post-processing after the prediction itself, the in-elastic strain is calculated from the incremental in-elastic strain via integration. Due to this in Figure 7, where the predictions quality for the (total) in-elastic strain is shown for the seven mission profiles, seven connected lines of prediction points are visible.

The application of integration also means the integration of errors. This has both advantages and disadvantages. Assume that only one error occurs at a specific point in time. Then, from this data point onwards, a deviation can be seen in all subsequent data points in Figure 7. This deviation will also be included in the MAPE calculation. On the other hand, a further, opposing error can cancel out the original error. Nevertheless, the prediction is sufficiently accurate, as also can be seen in Figure 8.

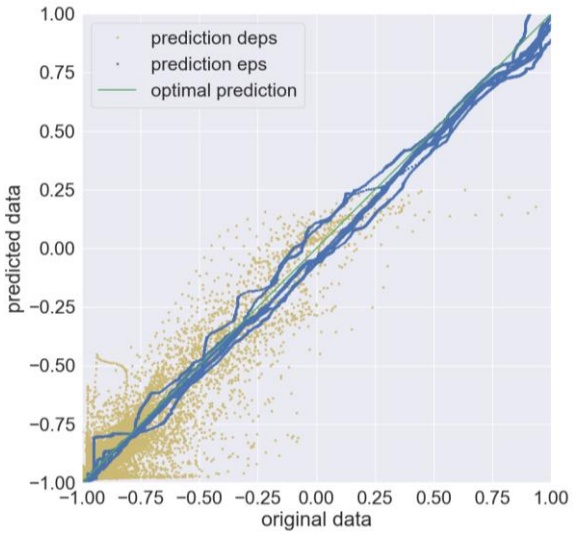


Figure 7: Prediction quality of the LSTM for in-elastic strain prediction. The yellow points indicate the prediction quality for the incremental strain, the blue for total strain.

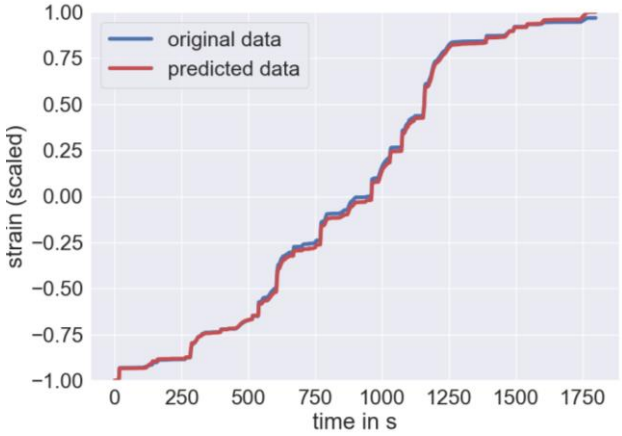


Figure 8: Comparison of the result of the FE simulation (blue) and the LSTM model prediction (red) for the WLTP cycle. The strain is scaled to (-1, 1).

So, with this surrogate model a reduction of the calculation time from two days to 19 seconds (7 seconds for the prediction of the temperature and 12 seconds for the prediction of the strain) for the WLTP cycle of 1800 seconds is achieved.

4. State-of-Health and Remaining Useful Life

As mentioned before for the FE simulations, the rainflow counting algorithm is applied in order to transfer the load profile and the resulting continuously increasing in-elastic strain into separate cycles. By using Coffin-Manson lifetime model (equation (1)), which describes the shape of the strain

Wöhler curve in the low cycle fatigue range, the number of cycles to failure $N_{f,i}$ for each cycle i is calculated. The inelastic strain $\Delta\varepsilon_{cr,i}$ for each sub-cycle per cycle (closed cycles within a larger cycle) is subtracted. From the number of cycles to failure the damage per cycle

$$D_i = 1/N_{f,i} \tag{2}$$

is calculated. Then Palmgren-Miner linear damage accumulation relation (proposed by Palmgren (1924) and further developed by Miner (1945)), is used to sum up all damage contributions

$$D = \sum_i D_i = \sum_i \frac{1}{N_{f,i}} \tag{3}$$

From this, the damage can be transferred into the state-of-health

$$SoH = 1 - D \tag{4}$$

Subsequently, the state of health was calculated for a series of WLTP cycles, whereby initially the WLTP cycles were not changed for the entire period. With this a lifetime of 143 days is calculated. In addition, when a health status of 50% was reached, the load level of the WLTP cycle was reduced. This is used to simulate a reduction in power in response to the damage reached. As shown in Figure 9, this reduction in power significantly increases the lifetime.

Consequently, this model can be used to adjust the load of the TO220 rectifier to the current state of health. Due to the short calculation time of the surrogate model, the influence of the load reduction on the lifetime can be predicted. This allows to adjust the load in a targeted manner, which is necessary for a control.

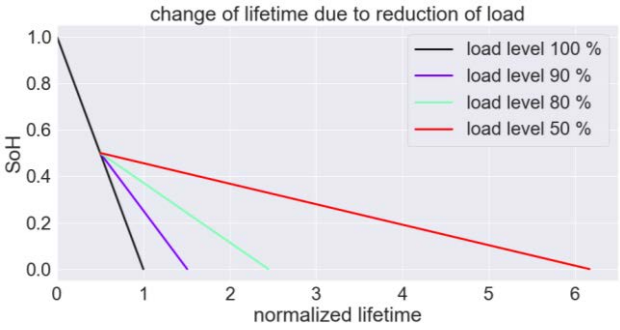


Figure 9: Change of the lifetime due to the reduction of the load level of the WLTP cycle.

5. Conclusion

In this work the methodology of replacing the FE simulation model by a surrogate model is shown. The amount of calculation time is massively reduced and due to this the surrogate model will be implemented on a micro controller in order to finalize the digital twin.

For the prediction of the temperature the final trained model is a LSTM model with just one unit layer and six units therein. Due to this low complexity, in future work a change to a less complicated model type will be taken under consideration.

Additionally, the surrogate model is currently being trained with a complete WLTP cycle. This has the disadvantage that the generation of the training data requires a relatively large amount of resources. However, it can be assumed that it contains multiple pieces of information that are not necessarily required for training the surrogate model. For this reason, the training of the surrogate model is to be simplified in future work. The WLTP cycle and other realistic load profiles will be analyzed using methods from time series analysis (e.g., the matrix profile, what is presented in Imani, Madrid, Ding, Crouter, and Keogh (2018) or Mercer, Alaee, Abdoli, Singh, Murillo, and Keogh (2021)), the relevant patterns and anomalies will be determined and calculated as separate profiles, weighted and specified in the training. This reduces the effort required to generate the training data.

Acknowledgement

This work was partially supported by HiEFFICIENT project. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement no. 101007281. The JU receives support from the European Union's Horizon 2020 research and innovation program and Austria, Germany, Slovenia, Netherlands, Belgium, Slovakia, France, Italy, and Turkey. Further, it was partially supported by Trust-E project. This project has received funding from PENTA - EUROPIDES[2] and BMBF (16ME0324).

Furtheron, Wolfgang Wondrak and Leonhard Hertenstein from Mercedes Benz AG are highly acknowledged for providing the current profile of the WLTP driven cycle.

References

Albrecht, J., Horn, T., Habenicht, S., & Rzepka, S. (2023, December). Mission Profile related Design for Reliability for Power Electronics based on Finite Element Simulation. In *2023 IEEE 25th Electronics Packaging Technology Conference (EPTC)* (pp. 722-726). IEEE. doi: 10.1109/EPTC59621.2023.10457766.

Chacko, J., Moeller, H., Kolas, K. A., Albrecht, J., & Rzepka, S. (2024, April). Investigation of an approach for the determination of the current State-of-Health and improvement of service life in power modules. In *2024 25th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE)* (pp. 1-9). IEEE. *doi: 10.1109/EuroSimE60745.2024.10491510*

Darveaux, R., & Banerji, K. (1991, May). Fatigue analysis of flip chip assemblies using thermal stress simulations and a Coffin-Manson relation. In *1991 Proceedings 41st*

Electronic Components & Technology Conference (pp. 797-805). IEEE. doi: 10.1109/ECTC.1991.163971

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780, doi: 10.1162/neco.1997.9.8.1735

Imani, S., Madrid, F., Ding, W., Crouter, S., & Keogh, E. (2018). Matrix profile XIII: Time series snippets: A new primitive for time series data mining. In *2018 IEEE international conference on big knowledge (ICBK)* (pp. 382-389). IEEE. doi: 10.1109/ICBK.2018.00058

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mercer, R., Alaee, S., Abdoli, A., Singh, S., Murillo, A., & Keogh, E. (2021). Matrix profile xxiii: Contrast profile: A novel time series primitive that allows real world classification. In *2021 IEEE International Conference on Data Mining (ICDM)* (pp. 1240-1245). IEEE. doi: 10.1109/ICDM51629.2021.00151

Miner, M. A. (1945). Cumulative Damage In Fatigue. *J. Appl. Mech.*, 12(3), A159-A164 Doi: https://doi.org/10.1115/1.4009458

Moeller, H., Inamdar, A., van Driel, W.D., Bredberg, J., Hille, P., Knoll, H. & Vandevelde, B. (2024). Digital Twin Technology in Electronics. In van Driel, W.D., Pressel, K. & Soytürk, M. (Eds.), *Recent Advances in the Reliability Assessment of Electronic Devices* (287-328). Location: Publisher. *(in press)*

Nexperia (2023). PSC1065 650 V, 10A SiC Schottky diode in TO-220-2 R2P. https://www.nexperia.com/products/diodes/silicon-carbide-sic-schottky-diodes/PSC1065K.html

Palmgren, A. (1924). Durability of ball bearings. *ZVDI, 68*(14), 339-341.

SAE International (2021) *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (SAE J3016).

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017, June). Long short-term memory network for remaining useful life estimation. In *2017 IEEE international conference on prognostics and health management (ICPHM) (pp. 88-95)*. IEEE.

Integrating Network Theory and SHAP Analysis for Enhanced RUL Prediction in Aeronautics

Yazan Alomari[1], Marcia L. Baptista [2], and Mátyás Andó[3]

[1,3]*Eötvös Loránd University, Faculty of Informatics, Institute of Computer Science, Budapest, 1053, Hungary*
yazan@inf.elte.hu
am@inf.elte.hu

[2]*Delft University of Technology, Faculty of Aerospace Engineering, GB Delft, 2600, The Netherlands*
m.lbaptista@tudelft.nl

ABSTRACT

The prediction of Remaining Useful Life (RUL) in aerospace engines is a challenge due to the complexity of these systems and the often-opaque nature of machine learning models. This opaqueness complicates the usability of predictions in scenarios where transparency is crucial for safety and operational decision-making. Our research introduces the machine learning framework that significantly improves both the **interpretability** and **accuracy** of RUL predictions. This framework incorporates SHapley Additive exPlanations (SHAP) with a surrogate model and Network Theory to clarify the decision-making processes in complex predictive models and enhance the understanding of the hidden pattern of features interaction. We developed a Feature Interaction Network (FIN) that uses SHAP values for node sizing and SHAP interaction values for edge weighting, offering detailed insights into the interdependencies among features that affect RUL predictions. Our approach was tested across 44 engines, showing RMSE values between 2 and 17 and NASA Scores from 0.2 to 1.5, indicating an increase in prediction accuracy. Furthermore, regarding interpretability the application of our FIN, revealed significant interactions among corrective speed and critical temperature points key factors in engine efficiency and performance.

1. INTRODUCTION

In the interdisciplinary domain of Prognostics and Health Management (PHM), the accurate prediction of Remaining Useful Life (RUL) for industrial assets has become paramount (Ren et al., 2023). As aerospace, automotive, and manufacturing sectors increasingly depend on the reliability of their machinery, accurately predicting maintenance needs has become essential for ensuring safety, maximizing efficiency, and reducing costs. This necessity has driven the shift from traditional prognostic methods to advanced machine learning techniques (Calabrese et al., 2020; Deutsch & He, 2018). These modern methods utilize large datasets to effectively identify complex patterns and trends in machinery wear and tear, significantly enhancing our ability to predict equipment failures (Duc Nguyen et al., 2019).

However, the application of ML in PHM is limited by significant challenges, such as models interpretability. The "black box" nature of many ML algorithms, particularly those based on deep learning, obscures the decision-making processes underlying their predictions. This opacity is a considerable concern in fields when understanding the 'why' behind a prediction is as critical as the prediction itself, necessitating models that stakeholders can trust and interpret (Baptista et al., 2022; Kononov et al., 2023; Vollert et al., 2021).

Historical reliance on reliability and physics based models for RUL estimation, though effective, often staggers upon the complexities inherent in real-world operational scenarios. These traditional methods necessitate detailed domain knowledge and often lack the flexibility to adapt to different types of machinery (X. Li et al., 2018; Si et al., 2011; Yan et al., 2021). The integration of machine learning into PHM, especially with the advent of sophisticated algorithms and the increased availability of sensor data opens a new opportunities in RUL prediction. This new technologies is characterized by learning from historical performance data, detecting subtle patterns, and predicting future outcomes with increased accuracy (A. Li et al., 2018; Yang et al., 2020).

The diversity and complexity of data in PHM, combined with the unique operational characteristics of different machinery, pose additional obstacles. These factors complicate the task of creating generalized models that are both accurate and interpretable across varied contexts (Lakkaraju et al., (2016). The need for models that can adapt to such diversity while

providing clear insights into their predictions is pressing (Rudin et al., 2022).

In response to these challenges, our research presents a framework that integrates Network Theory with SHapley Additive exPlanations (SHAP) to enhance the accuracy and the interpretability of ML-based RUL predictions. Our contributions are manifold and aim to bridge the gap between complexity and explainability:

I. **Application of a SHAP Theory**: We use a surrogate Model to translate complex interactions from deep learning models into SHAP interaction values.

II. **Application of Network Theory and FIN**: The SHAP interaction values are mapped onto the Feature Interaction Network. This mapping provides a detailed analysis of feature interdependencies to improve our understanding of the factors that affect the reliability of machinery.

III. **Integration of SHAP and Network Theory**: The proposed method combines SHAP values with Network Theory to develop an "augmented" Feature Interaction Network (FIN). This network helps clarify and quantify how different features interact and influence RUL predictions.

The novelty of this research lies in the fusion of network theory with feature importance methodologies to decode the nuanced interplay of operational parameters. By applying a Feature Interaction Network (FIN), a structural map of feature interdependencies further enriched by the integration of SHAP values, we were quantifying and explain feature contributions. Central to this approach was the novel application of surrogate models, facilitating the distillation of SHAP interaction effects into discernible edge strengths within the FIN. Concurrently, the combination of mean absolute SHAP values with network centrality metrics allows positioning a more comprehensive description of feature significance and influence. This research aims to envelope an innovative yet pragmatic set of tools, that can enhance Explainability and interpretability of predictive maintenance practices.

The paper is organized as follows: Section II surveys related literature, establishing the context for our contributions. Section III details our methodology, highlighting the synergistic use of SHAP analysis and Network Theory to decode ML model decisions. Section IV discusses the empirical findings, focusing on the insights gleaned from the FIN and its practical implications for PHM. Section V concludes, reflecting on the impact of our work and suggesting directions for future research in enhancing model transparency and reliability.

2. BACKGROUND AND LITERATURE REVIEW

In the field of prognostics and health management (PHM), the ability to accurately predict the Remaining Useful Life (RUL) of machinery is gaining traction (Lei et al., 2018; Ramezani et al., 2019; Zhao & Addepalli, 2020). This increased popularity is largely driven by advancements in machine learning and deep learning technologies. (Berghout & Benbouzid, 2022; Chen, Wu, Zhao, Guretno, Yan, Member, et al., 2021; Ferreira & Gonçalves, 2022). This review aims to summarize recent developments in RUL prediction, highlighting the evolution of methodologies and techniques across various industrial sectors.

The significant improvements in RUL prediction began with the innovative preprocessing of sensor data. For instance, Ensarioğlu et al., (2023) introduced a method that combined difference-based feature construction with a hybrid 1D-CNN-LSTM model, enhancing prediction accuracy significantly. Among the more notable preprocessing techniques is the sliding time window method, which organizes time-series signals into segments of equal length for more consistent input data (Guo et al., 2022). While effective, this method can be labor-intensive and somewhat dependent on the operator's expertise. Another valuable technique is the short-time Fourier transform (STFT), which considers the time correlation of signal sequences, providing a robust basis for subsequent analyses (Liu et al., 2022; Zhang et al., 2023). Also, the integration of long and short-term memory networks (LSTMs) with convolutional block attention modules has improved our understanding of neural decision-making processes (Remadna et al., 2023). The application of deep convolutional variational autoencoders equipped with attention mechanisms has improved the spatial distribution of features, thereby enhancing the interpretability of predictive models (Cheng et al., 2022).

The interpretability of machine learning techniques in RUL prediction has seen significant advancements, particularly through the integration of attention mechanisms and feature fusion frameworks. An attention-based deep learning framework was developed to effectively combine handcrafted and automated features for accurate RUL prediction, demonstrating high efficiency performance on real datasets (Chen, Wu, Zhao, Guretno, Yan, & Li, 2021). Remadna et al., (2023) proposed a fusion of an attention-based convolutional variational autoencoder with an ensemble learning classifier, achieving high accuracy and improved interpretability. Watson (2020) highlighted the conceptual challenges in interpretable machine learning (IML), emphasizing the need for clarity in target definitions and the importance of error rate considerations and testing for IML algorithms. Additionally, Xu et al., (2022) introduced an approaches combined deep learning with other techniques such as particle filters and knowledge distillation to enhance feature extraction, interpretability, and model compression for efficient RUL prediction.

Ye & Yu, (2023) introduced the Selective Adversarial Adaptation Network (SAAN), an approach to domain adaptation employing selective feature interaction for effective knowledge transfer in machine RUL prediction under variable conditions. Kobayashi et al., (2023), also highlighted the critical need for transparency and interpretability in AI models, emphasizing the significance of Explainable AI (XAI) and Interpretable Machine Learning (IML) in RUL prediction in digital twin systems. Zou et al., (2021) proposed an approach for RUL prediction in small data scenarios using a fully convolutional variational auto-encoding network, effectively addressing underfitting issues and demonstrating superior performance in degradation feature extraction and failure threshold determination compared to traditional models.

LIME was proposed by(Ribeiro et al., 2016) as a local model-agnostic approach to interpretability. It has been since then used extensively in prognostics and health management. LIME is a local-model because it approximates the learning model with an interpretable simplified surrogate around a single prediction. As a model0agnostic approach, LIME is a generic and works with any underlying predictive model.

This method has been particularly useful for RUL prediction, as it allows engineers to understand the impact of different features on the predicted outcomes. For instance, Khan et al., (2022); Serradilla Oscar et al., (2020) demonstrated the efficacy of LIME in explaining RUL predictions, enabling a deeper understanding of the degradation patterns and contributing factors, thus facilitating more informed maintenance decisions.

In a recent study, Alomari et al., (2023)developed a comprehensive method for predicting the Remaining Useful Life (RUL) of aircraft engines. Our approach integrates advanced feature engineering, dimensionality reduction through principal component analysis, and a range of feature selection techniques, including Genetic Algorithms, Recursive Feature Elimination, Least Absolute Shrinkage and Selection Operator Regression, and Feature Importances from Random Forest models. A significant innovation in this research is the introduction of the Aggregated Feature Importances with Cross-validation (AFICv) technique. This method enhances the selection process by prioritizing features based on their mean importance also establishes a selection criterion that retains features contributing up to 70% of the cumulative mean sum which is effectively simplifies the model complexity. Another finding in our research is introducing a novel PCA-based interpretability framework to provide actionable insights and enhance the practical utility of our findings for domain experts in the aerospace industry.

2.1. Data Description

The N-CMAPSS dataset (Chao et al., 2021) is a dataset that uses real flight conditions from a commercial jet to simulate the operative conditions (w) within its model. This dataset provides synthetic degradation trajectories for a fleet of turbofan engines, effectively replicating various unknown initial health states under authentic flight conditions. It includes eight distinct datasets derived from 128 engines, each illustrating seven unique failure modes. These modes predominantly affect the flow (F) and efficiency (E) of key engine components such as the fan, low-pressure compressor (LPC), high-pressure compressor (HPC), high-pressure turbine (HPT), and low-pressure turbine (LPT).

Flight conditions within the N-CMAPSS model are categorized into three distinct classes based on the length of the flight. The details of these flight classes, along with the specific failure modes for each dataset, are meticulously documented in Table 1 (4 datasets were used only from the entire original dataset). Additionally, the dataset provides extensive information on the scenario descriptors as in Table 1, and measurements and virtual sensors, which are thoroughly described in the turbofan Jet engine schematic representation Figure 1 and Table 2. This structured approach in modeling the failure modes and operational conditions forms the backbone of the current model development, offering a realistic and detailed perspective of engine degradation under varied flight scenarios.

Table 1 N-CMAPSS Datasets overview

Name	# Units	Flight Classes	Failure Modes
DS01	10	[1 - 2 - 3]	1
DS02	9	[1 - 2 - 3]	2
DS03	15	[1 - 2 - 3]	1
DS05	10	[1 - 2 - 3]	1

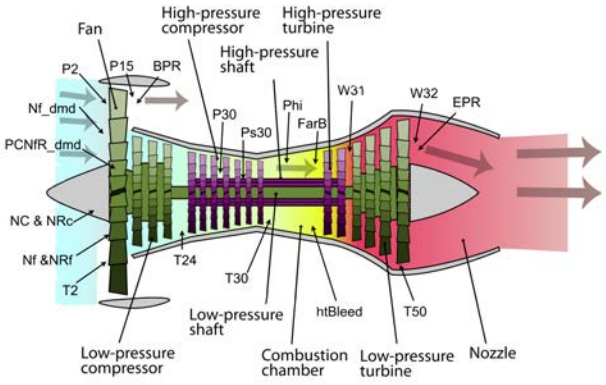


Figure 1 Turbofan Jet Engine Schematic Representation

2.2. Data preprocessing and feature selection

Standardization was applied to the dataset, as detailed in equations 1-3, normalizing each feature to have zero mean and unit variance. This step was essential for ensuring consistency across different data scales and enhancing the efficacy of the subsequent feature selection and machine learning models:

$$Standardization \quad Z = \frac{x - \mu}{\sigma} \tag{1}$$

$$Mean \quad \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{2}$$

$$Standard\ Deviation \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

Feature selection is critical in prognostics and health management, offering substantial benefits for applications such as RUL prediction and fault detection. By eliminating redundant features, this process effectively reduces the input dimensions for machine learning models, thereby enhancing their performance by focusing on the most informative attributes (Alomari et al., 2023; Aremu et al., 2020).

In this study, features are selected based on their statistical variability. Sensors that exhibit zero standard deviation, indicating no variation and thus no predictive value, have been excluded. An example of this selection process can be seen in Figure 2, which illustrates how sensors are chosen based on their variability over time. In Figure 2, features such as 'T2', 'W50', and 'Nc' exhibit fluctuating values, whereas other features remain constant, indicating they provide limited informational value to the mode.

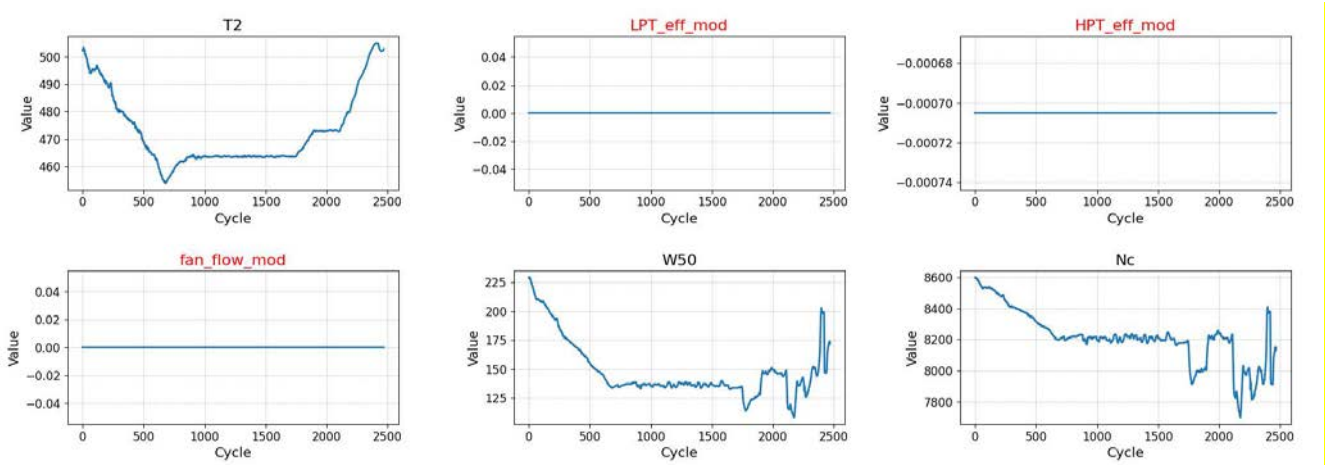


Figure 2 Variability Analysis of Sensor Data for Feature Selection

This selection was specifically tailored to exclude sensors with negligible fluctuations or redundant information, focusing instead on those providing significant insights into engine performance and wear. The final selected features are listed in Table 2.

Table 2 list of the selected features

alt	Mach	TRA	T2	T24
P24	Ps30	P40	P50	Nf
T30	T48	T50	P15	P2
Nc	Wf	T40	P30	P21

3. METHODOLOGY

The methodology, illustrated in Figure 3, is based on a composite model integrating Deep Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), a customized Time Distributed Attention mechanism, and an innovative Feature Interaction Network (FIN). The goals are to improved the precision and interpretability of RUL predictions for aircraft engines.

The GRU layers illustrated in Figure 4 capture the temporal correlations within the sequential engine data, while the CNN layers distill critical spatial features, thereby enhancing the model's capability to identify salient patterns indicative of engine failure. The custom attention layer defined in Figure 5 allows to selectively simplify temporal events within the engine's operational history, further refining the model's predictive accuracy.

To enhance interpretability, the FIN, constructed using SHAP (SHapley Additive exPlanations) values shown in Figure 10, quantifies the impact and interactions of individual features. The node's size within the FIN is representative of the mean absolute SHAP values. This allows better demonstration the feature importance visually. Edge weights are defined by SHAP interaction values, illustrating the strength of the interaction between each pair of features.

The methodological combination of GRU, CNN, attention mechanisms, and SHAP-driven FIN proffers a multidimensional interpretable approach, in the field of aerospace prognostics and health management.

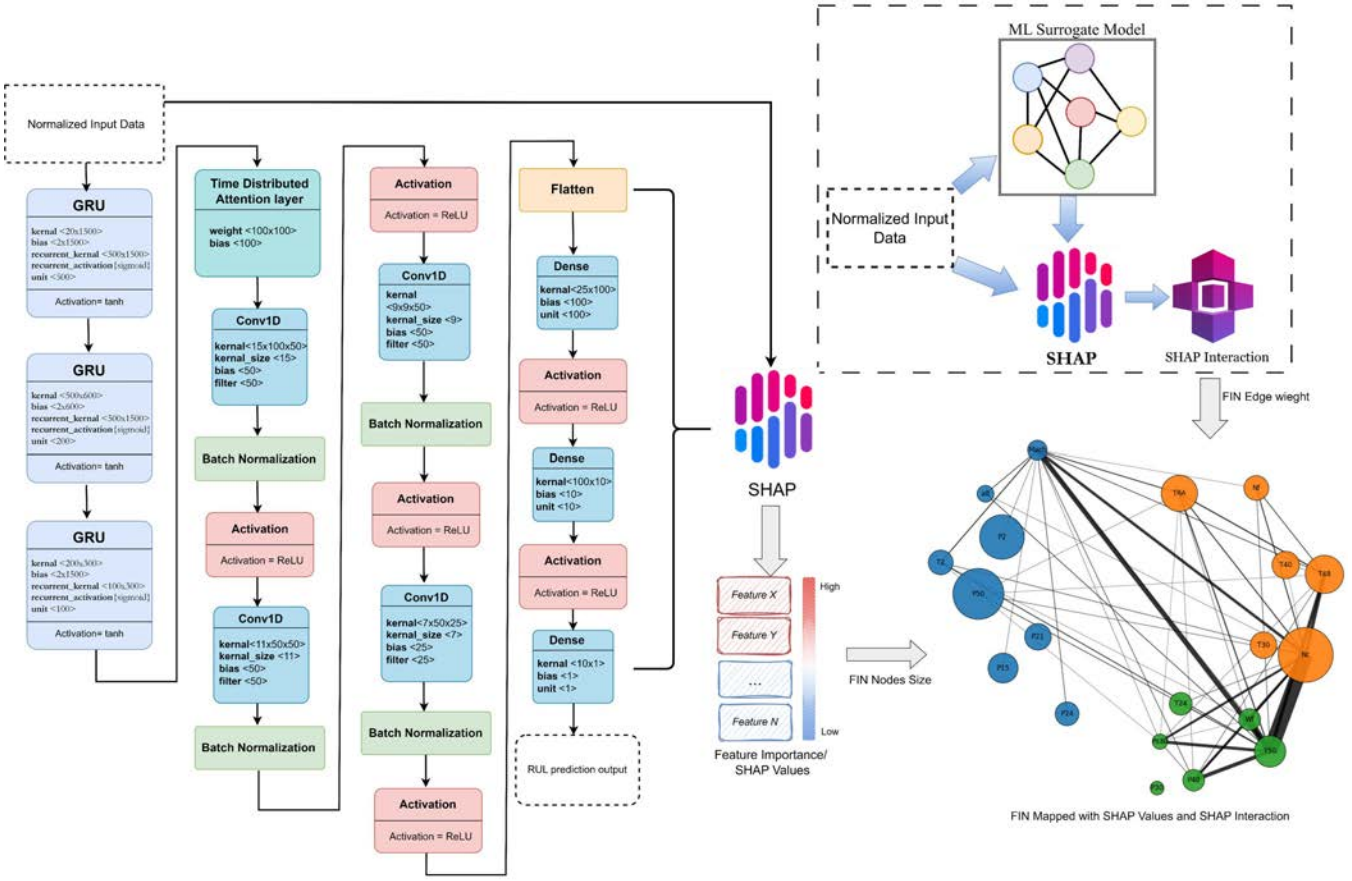


Figure 3 proposed model for RUL prediction and FIN

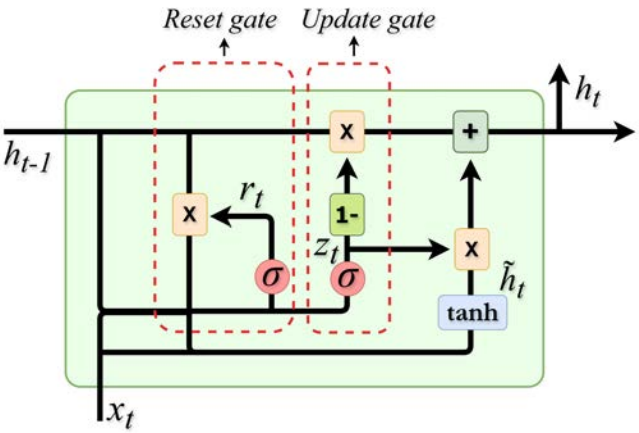


Figure 4 Gated recurrent unit (GRU) neural network structure

The GRU, presented in Figure 4, introduced by (Cho et al., 2014), is a type of recurrent neural network designed to model temporal sequences and long-range dependencies more effectively than standard RNNs. They simplify the recurrent module while retaining the ability to capture dependencies in time-series data, making them computationally efficient and powerful for tasks such as speech recognition, language modeling, and sequential prediction, which are crucial in PHM contexts (Cao et al., 2021; Zhou et al., 2022, Zhou et al., 2023). The core functionality of GRUs relies on the modulation of information flow across sequence steps, controlled by the update and reset gates.

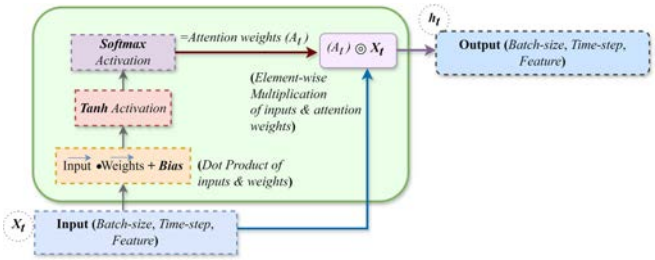


Figure 5 time-distributed attention mechanism

After the GRU layers, a customized Time Distributed Attention mechanism (see Figure 5) was used to improve the model's ability to focus on the most critical features within the sequential data. It applies an attention mechanism to each time step independently. This is achieved by computing an attention score for each feature using a learned weight matrix and bias vector. The scores are then normalized via a softmax function to create attention weights, which are subsequently used to scale the input features. This process allows the model to dynamically prioritize significant information, thereby improving the interpretability and accuracy of RUL predictions.

3.1. SHapley Additive exPlanations (SHAP)

In the field of explainable artificial intelligence (XAI), Shapley Additive exPlanations (SHAP) (Lundberg et al., 2017) values are a central tool for quantifying the contributions of individual features to a model's prediction. Rooted in cooperative game theory, SHAP values, formally described in Equation (4), enable the measurement of each feature's influence by comparing the model's output with and without the presence of the feature. This approach not only fosters transparency but also imbues the analysis with a rigorous mathematical foundation.

SHAP is crucial to PHM (Alomari & Andó, 2024) where understanding the impact of various features on the prediction of system failures or maintenance needs is paramount. SHAP values facilitate this by attributing precise, quantifiable contributions of individual features to the overall prediction of system health, thereby enabling more accurate and timely decision-making.

$$\text{SHAP}(j) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)] \quad (4)$$

S - is a subset of features
N - is the set of all features
$|S|$ - denotes the cardinality (size) of set $\boldsymbol{S}$
$f(S \cup \{j\})$ - is the prediction with both the features in set $\boldsymbol{S}$ and feature $\boldsymbol{j}$
$f(S)$ - is the prediction with just the features in set $\boldsymbol{S}$

Equation (4) computes the contribution of feature j by iterating over all possible subsets S of the remaining features in N and comparing the difference in the prediction when feature j is included versus when it is excluded.

3.2. Network Theory

Network Theory (Borgatti & Halgin, 2011) provides a framework for understanding the structure and dynamics of complex systems by visualizing them as networks of nodes (features) and edges (interactions). This approach is especially useful in PHM, to reveal the complex interdependencies between system components. By applying Network Theory to create a Feature Interaction Network (FIN), we can perform both visual and quantitative analyses of how individual system features interact and collectively impact overall system behavior (see Figure 6). The decision to use a FIN was deliberate; it helps in mapping out the relationships and dependencies among features effectively and also simplifies the understanding of complex data structures for engineers and domain experts.

To accurately model the interactions within a FIN, the Graphical Lasso (GLasso) algorithm (Friedman et al., 2008) was utilized. GLasso effectively determines the conditional independence structure between variables (features), offering a sparse representation of the feature interaction network. The mathematical formulation of Glasso (Equation 5) is centered on optimizing the following objective function:

$$\min_{\Theta} - \log\det(\Theta) + tr(\mathcal{S}\Theta) + \lambda\|\Theta\|_1 \quad (5)$$

Here, Θ represents the precision matrix (inverse covariance matrix) to be estimated, $\mathcal{S}$ is the empirical covariance matrix of the data, $log\ \det(\Theta)$ ensures the positive definiteness of Θ, $tr(\mathcal{S}\Theta)$ is the trace term encouraging fidelity to the observed data, $\|\Theta\|_1$ denotes the L_1 norm imposing sparsity, and λ is a regularization parameter controlling the degree of sparsity. By solving this optimization problem, GLasso identifies significant interactions while discarding the insignificant, resulting in a FIN that highlights the most crucial feature relationships.

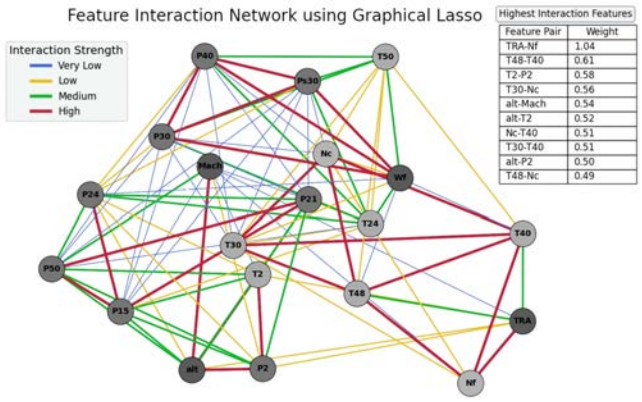


Figure 6 Basic FIN graph with interaction strength

Within the framework of Network Theory, centrality measures serve as tools to compute the prominence of individual nodes. The centrality measures, namely degree centrality for direct linkages, betweenness centrality for intermediary influence, and closeness centrality for overall accessibility are presented in equations (6-8), are instrumental in discerning the structural backbone of the Feature Interaction Network.

Degree centrality $(\boldsymbol{C_D})$ of a node v is defined as the fraction of nodes it is connected to. It reflects the immediate influence of a node within the network.

$$C_D(v) = \frac{\deg(v)}{N-1} \tag{6}$$

Where $\deg(v)$ is the degree of node v (i.e., the number of edges incident to v and N is the total number of nodes in the network. Degree centrality helps us pinpoint features that exert considerable control over the system's behavior, thereby identifying potential points of proactive maintenance and intervention.

Betweenness centrality $(\boldsymbol{C_B})$ of a node v quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

$$C_B(v) = \sum_{\mathcal{S} \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7}$$

Where σ_{st} is the total number of shortest paths from node $\mathcal{S}$ to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v, it highlights the node's role in facilitating interactions between other nodes. Identifying such nodes helps in strategizing interventions that can prevent cascading failures in engine operations.

Closeness centrality $\boldsymbol{C_c(v)}$ measures how close a node is to all other nodes in the network, indicating how easily information can flow from the given node to others.

$$C_c(v) = \frac{N-1}{\sum_{u \neq v} d(v,u)} \tag{8}$$

Where $d(v,u)$ is the shortest path distance between nodes v and u, and N is the total number of nodes in the network. Features with high closeness centrality are likely to affect the system more rapidly, making them critical targets for monitoring and early preventive maintenance.

Community detection algorithms, such as the Louvain method (De Meo et al., 2011) given by equation (9), partition the network into communities or clusters of nodes that are more densely connected with each other than with the rest of the network. This segmentation can reveal modular structures within the feature set, suggesting subsystems within the engine that have distinct behaviors.

$$Q = \frac{1}{2M} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{9}$$

The following definitions characterize the parameters of the algorithms

- A_{ij} represents the weight of the edge between nodes i and j. For unweighted networks, A_{ij} is 1 if there is an edge between i and j, and 0 otherwise.
- $k_i \; and \; k_j$ are the sum of the weights of the edges attached to nodes i and j, respectively.
- m is the sum of all the edge weights in the network.
- $c_i \; and \; c_j$ are the communities of nodes i and j
- δ is the Kronecker delta function, which is 1 if $c_i = c_j$ (i.e., nodes i and j are in the same community) and 0 otherwise.

The goal of the Louvain method is to maximize Q through a heuristic approach that iteratively groups nodes into communities.

3.3. Evaluation metrics

The proposed model evaluation was conducted using the N-CMAPSS datasets (DS01, DS02, DS03 and DS05), focusing on the accuracy of Remaining Useful Life (RUL) predictions. The performance of our proposed model was primarily assessed by measuring the discrepancy between the predicted and actual RUL values. For this purpose, we employed two key metrics: Root Mean Square Error (RMSE) and NASA Score (Saxena et al., 2008), as defined in equations (10 - 13). These metrics, calculated over the number of data points (n), provided a comprehensive understanding of the model's

predictive accuracy and reliability in various operational scenarios presented within the N-CMAPSS datasets.

$$RMSE\ (P_{RUL}, T_{RUL}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_{RUL} - T_{RUL})^2} \quad (10)$$

$$\text{NASA Score}_i = \begin{cases} exp\left(-\frac{\Delta_i}{10}\right) - if\ \Delta_i < 0 \\ exp\left(-\frac{\Delta_i}{13}\right) - if\ \Delta_i \geq 0 \end{cases} \quad (11)$$

Where:

$$\Delta_i = RUL_{predicted\ i} - RUL_{true_i} \quad (12)$$

$$NASA\ Score = \frac{1}{n}\sum_{i=1}^{n} NASA\ Score_i \quad (13)$$

4. Results and Discussion

Our enhanced approach to predicting Remaining Useful Life (RUL) uses an integration of Deep Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), and a custom Time Distributed Attention mechanism. This tailored combination has advanced the accuracy of RUL predictions by effectively capturing complex temporal and spatial patterns within engine operational data, crucial for early and accurate fault detection. The inclusion of the Custom Attention Layer allows for identifying critical features and time steps, significantly refining the interpretability of our predictive models. The effectiveness of these innovations is substantiated by our empirical results presented in Tables 3 and 4. Table 3 includes a comparison of some of our results with three methods from the literature, while Table 4 presents the results for the remaining engines, for which no direct comparisons to existing studies could be made.

Across the different datasets, the model demonstrates proficiency in RUL prediction, as evidenced by the calculated Root Mean Square Error (RMSE) and the NASA prognostics score, with a significant performance in the critical RUL phase. This is important since the latter half of life where accurate prediction is most vital. Particularly significant are the outcomes on DS02 and DS03, where the model achieves RMSE values as low as 2 cycles for the critical RUL, alongside correspondingly low NASA-scores, highlighting the model's precision in the most consequential phase of the engine's lifecycle.

The visualization of the RUL prediction and critical RUL of two engines, 9 and 12, from DS01 and DS03, respectively, along with their SHAP values, is presented in Figures 7 and 8. These figures illustrate the model's ability to track the Remaining Useful Life (RUL) over engine cycles accurately, with a particular focus on the critical RUL phase. The SHAP interpretation plots highlight the influence of various sensors on the model's predictions.

For Engine 12, significant features include 'Nc' (corrective speed), 'P50' (pressure at the fan outlet), and 'T2' (temperature at the fan inlet). The high SHAP values for these features indicate their substantial impact on the RUL predictions. Specifically, 'Nc' demonstrates a strong correlation with the engine's operational efficiency, reflecting its role in adaptive speed control. Similarly, 'P50' and 'T2' provide crucial insights into the pressure and temperature dynamics, essential for accurate prognostics.

In Engine 9, the SHAP values reveal 'Nc', 'T50' (temperature at the engine outlet), and 'P2' (pressure at the fan inlet) as key contributors. The interactions between 'Nc' and 'T50' (as they have opposite influence) suggest that the corrective speed adjustments are heavily influenced by thermal conditions at critical engine points. The significant SHAP values for 'P2' underscore the importance of pressure measurements in anticipating engine failures.

Table 3 Prognostics performance assessment comparison with different methods

Dataset	**Engine**	RMSE Proposed	RMSE Literature (Koutroulis et al., 2022)	RMSE Literature LR+ (Maulana et al., 2023)	RMSE Literature MLP+ (Maulana et al., 2023)
DS02	11	4.7	5.1	11.4	11.5
	14	6.1	11.9	10.9	11.1
	15	4	5.8	8.9	18.2
DS03	13	3.9	6.8	--	--
	14	3.2	5.1	--	--
	15	2.1	3.04	--	--

Table 4 Comparative assessment of RMSE and NASA-Score metrics for RUL prediction across engine units

Dataset	Engine	RMSE	RMSE Critical RUL	NASA-Score	NASA-Score Critical RUL
DS01	**7**	8.4	7	1.1	0.9
	8	6	4	0.6	0.5
	9	14	12	2.1	1.5
	10	5	3.5	0.5	0.37
DS03	**10**	8	2.9	0.8	0.22
	11	8	3	0.8	0.25
	12	17	7.3	6.4	1
DS05	**7**	10	3.8	1.9	0.37
	8	6	2.4	0.7	0.2
	9	7	3.6	0.8	0.27
	10	9	3.8	1.2	0.28

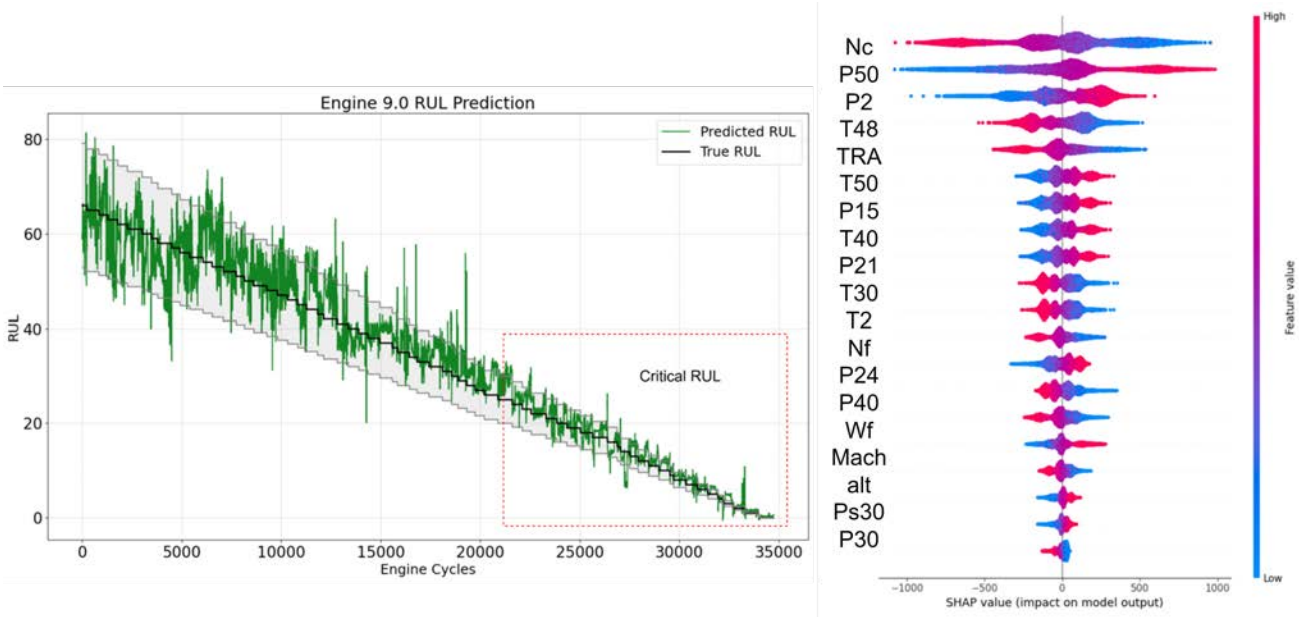


Figure 3 RUL prediction of engine 9 of DS01 with SHAP summary

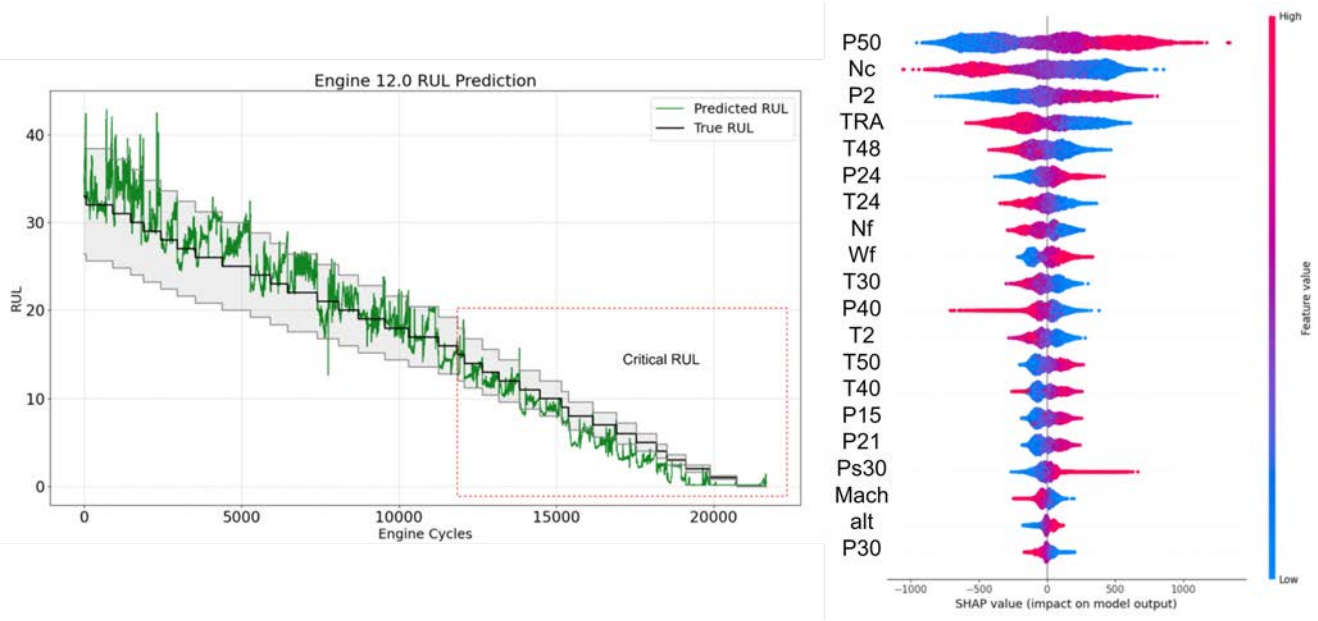


Figure 4 RUL prediction of engine 12 of DS03 with SHAP summary

The Feature Interaction Network (FIN) in Figure 8 provides an overview of the complex relationships inherent in the proposed predictive model for Remaining Useful Life (RUL). Through community detection algorithms, it has discerned distinct clusters within the network, indicative of underlying structures where subsets of features exhibit tightly knit interactions, potentially alluding to functional modules within the engine's operational parameters. The community color-coding allows to observe the modular nature of feature interdependencies, which may correspond to different physical or operational aspects of engine performance. Additionally, the betweenness centrality analysis reveals key nodes such as 'TRA,' 'P24,' and 'P15' that act as critical conduits in the flow of information through the network, signifying their roles in the model's inference processes.

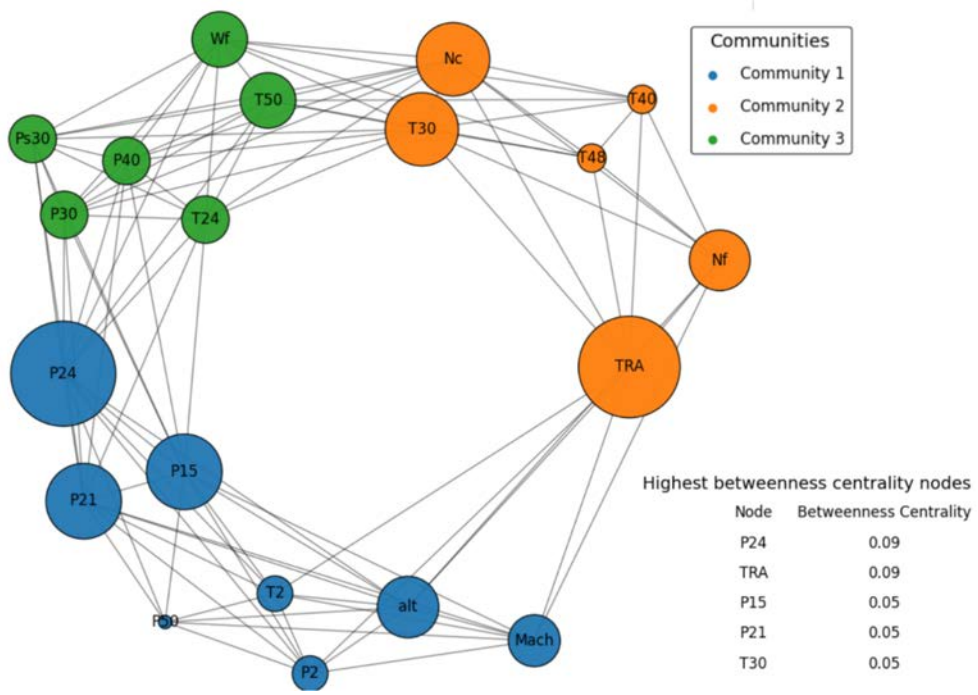


Figure 5 FIN with betweenness centrality and community detection

In Figure 10, we illustrate an innovative Feature Interaction Network (FIN) that leverages SHAP values to illuminate the complex dynamics within our predictive model for Remaining Useful Life (RUL) of aerospace engines. This network diagram, empowered by a surrogate model, not only visualizes the relative influence of various engine features but also clarifies their interrelationships. Each node, scaled according to mean absolute SHAP values, reflects the magnitude of influence each feature holds over the RUL predictions, with larger nodes marking more influential features.

These nodes are distinctly color-coded to represent different communities or clusters of features that share similar behavior patterns within the predictive framework, highlighting how groups of related features collectively impact engine performance. The edges between nodes, whose thickness is determined by the SHAP interaction values, illustrate the strength of the interactions between feature pairs, revealing critical dependencies and synergies.

Key interactions such as those between 'NC' (corrective speed) and temperatures at critical engine locations ('T50' and 'T48') suggest a profound connection between engine speed adjustments and thermal conditions. This relationship is crucial for maintaining optimal engine performance, particularly under varying operational stresses. The interaction between 'NC' and 'T50' highlights how adjustments in engine speed can be crucial in managing the engine's thermal output to avoid overheating while maintaining efficiency.

Further, the interaction between 'T50' and 'Mach' (aircraft velocity relative to the speed of sound) underscores the significant impact of aerodynamic performance on engine thermal management. The relationship between engine thermal outputs and flight speed suggests that higher speeds may require adjustments in thermal management strategies to maintain engine integrity and performance.

Additionally, the 'NC - Mach' interaction points to a dynamic balancing act required between engine speed and aircraft velocity, indicating that engine control systems need to be highly adaptive to changes in flight dynamics. This adaptiveness is crucial for optimizing fuel consumption and minimizing wear and tear under different flight conditions.

Lastly, the interaction between 'T50' and 'P40' (pressure at the fan outlet) sheds light on how temperature and pressure management are interlinked, playing a pivotal role in ensuring the engine's thrust efficiency and overall stability. This insight is particularly valuable for developing more effective predictive maintenance strategies, aiming to reduce unexpected downtimes and extend the engine's useful life.

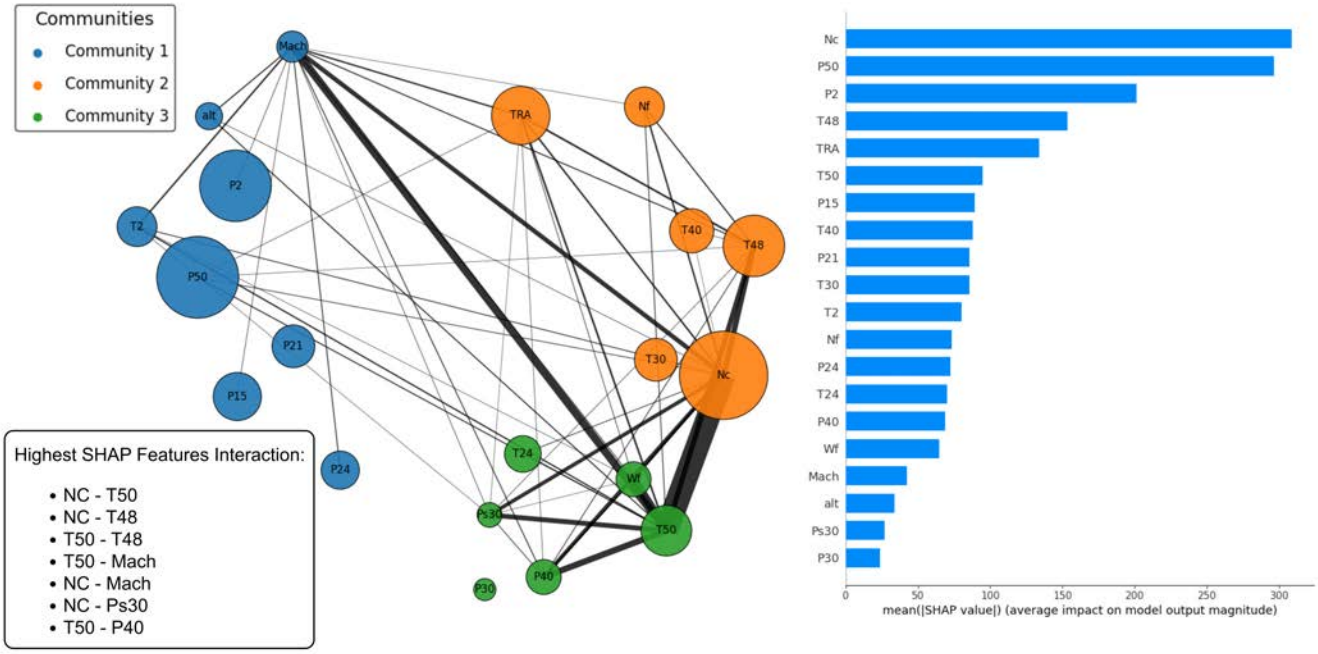


Figure 6 Feature Interaction Network (FIN) Visualizing Key Dependencies and Community Structures: This network map illustrates the Feature Interaction Network (FIN) with nodes sized according to their mean absolute SHAP values, which reflect the impact magnitude on the model's output. The nodes are color-coded by community, identifying clusters of tightly interconnected features that influence system behavior in distinct ways. Thicker lines between nodes indicate stronger SHAP

feature interactions, highlighting critical dependencies such as NC-T50 and T50-Mach, which are pivotal for understanding complex dynamics within the aerospace engine's operations

5. CONCLUSION

In conclusion, our research aims to advance the predictive maintenance field by developing a prognostic framework that combines cutting-edge machine learning techniques with innovative interpretative methodologies to predict the Remaining Useful Life (RUL) of aerospace engines. Utilizing a Surrogate Model, we have successfully mapped complex SHAP feature interactions into a well-defined Feature Interaction Network (FIN). This network, structured with nodes proportionally scaled by mean absolute SHAP values and connections defined by the strength of SHAP interactions, vividly represents the intricate relationships between operational parameters.

Our detailed analysis highlighted crucial feature interactions, notably between corrective speed and critical engine temperature, which are point factors essential for optimizing engine efficiency and performance. Furthermore, the application of community detection in the FIN has significantly deepened our understanding of these features, grouping related variables to illuminate how they collectively impact RUL predictions. This clustering clarifies the predictive model's structure and enhances the interpretability of the data, providing clear pathways for intervention.

The visual representation of the FIN is not merely an analytical tool; it acts as a vital conduit translating complex, data-driven insights into tangible, operational strategies. This visualization underscores the transformative potential of interpretative machine learning to convert abstract data into actionable intelligence, a resource of value in the high-stakes field of aerospace prognostics where the accuracy of predictions can directly influence operational safety and maintenance efficiency.

ACKNOWLEDGMENT

Project no. TKP2021-NVA-29 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme

REFERENCES

Alomari, Y., & Andó, M. (2024). SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis. *Results in Engineering*, *21*. https://doi.org/10.1016/j.rineng.2024.101834

Alomari, Y., Andó, M., & Baptista, M. L. (2023a). Advancing aircraft engine RUL predictions: an interpretable integrated approach of feature engineering and aggregated feature importance. *Scientific Reports 2023 13:1*, *13*(1), 1–14. https://doi.org/10.1038/s41598-023-40315-1

Alomari, Y., Andó, M., & Baptista, M. L. (2023b). Advancing aircraft engine RUL predictions: an interpretable integrated approach of feature engineering and aggregated feature importance. *Scientific Reports 2023 13:1*, *13*(1), 1–14. https://doi.org/10.1038/s41598-023-40315-1

Aremu, O. O., Cody, R. A., Hyland-Wood, D., & McAree, P. R. (2020). A relative entropy based feature selection framework for asset data in predictive maintenance. *Computers & Industrial Engineering*, *145*, 106536. https://doi.org/10.1016/J.CIE.2020.106536

Baptista, M. L., Goebel, K., & Henriques, E. M. P. (2022). Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artificial Intelligence*, *306*, 103667. https://doi.org/10.1016/J.ARTINT.2022.103667

Berghout, T., & Benbouzid, M. (2022). A Systematic Guide for Predicting Remaining Useful Life with Machine Learning. *Electronics 2022, Vol. 11, Page 1125*, *11*(7), 1125. https://doi.org/10.3390/ELECTRONICS11071125

Borgatti, S. P., & Halgin, D. S. (2011). On Network Theory. *Organization Science*, *22*(5), 1168–1181. https://doi.org/10.1287/orsc.1100.0641

Calabrese, F., Regattieri, A., Botti, L., Mora, C., & Galizia, F. G. (2020). Unsupervised fault detection and prediction of remaining useful life for online prognostic health management of mechanical systems. *Applied Sciences (Switzerland)*, *10*(12), 4120. https://doi.org/10.3390/APP10124120

Cao, Y., Jia, M., Ding, P., & Ding, Y. (2021). Transfer learning for remaining useful life prediction of multi-conditions bearings based on bidirectional-GRU network. *Measurement*, *178*, 109287. https://doi.org/10.1016/J.MEASUREMENT.2021.109287

Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *NASA Ames Research Center, Moffett Field*.

Chen, Z., Wu, M., Zhao, R., Guretno, F., Yan, R., & Li, X. (2021). Machine Remaining Useful Life Prediction via an Attention-Based Deep Learning Approach. *IEEE Transactions on Industrial Electronics*, *68*(3), 2521–2531. https://doi.org/10.1109/TIE.2020.2972443

Chen, Z., Wu, M., Zhao, R., Guretno, F., Yan, R., Member, S., & Li, X. (2021). Machine Remaining Useful Life

Prediction via an Attention-Based Deep Learning Approach. *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS*, *68*(3). https://doi.org/10.1109/TIE.2020.2972443

Cheng, Y., Hu, K., Wu, J., Zhu, H., & Shao, X. (2022). Autoencoder Quasi-Recurrent Neural Networks for Remaining Useful Life Prediction of Engineering Systems. *IEEE/ASME Transactions on Mechatronics*, *27*(2), 1081–1092. https://doi.org/10.1109/TMECH.2021.3079729

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv Preprint ArXiv*.

De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). *Generalized Louvain Method for Community Detection in Large Networks*. https://doi.org/10.1109/ISDA.2011.6121636

Deutsch, J., & He, D. (2018). Using deep learning-based approach to predict remaining useful life of rotating components. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*(1), 11–20. https://doi.org/10.1109/TSMC.2017.2697842

Duc Nguyen, V., Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., & Bäck, T. (2019). A Review: Prognostics and Health Management in Automotive and Aerospace. *International Journal of Prognostics and Health Management*, *10*(2), 35. https://www.klm.com/corporate/en/publications/2015

Ensarioğlu, K., İnkaya, T., & Emel, E. (2023). Remaining Useful Life Estimation of Turbofan Engines with Deep Learning Using Change-Point Detection Based Labeling and Feature Engineering. *Applied Sciences 2023, Vol. 13, Page 11893*, *13*(21), 11893. https://doi.org/10.3390/APP132111893

Ferreira, C., & Gonçalves, G. (2022). Remaining Useful Life prediction and challenges: A literature review on the use of Machine Learning Methods. *Journal of Manufacturing Systems*, *63*, 550–562. https://doi.org/10.1016/J.JMSY.2022.05.010

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Guo, R., Li, H., Huang, C., Zhao, C., Huang, X., Liu, H., Li, J., & Chen, R. (2022). A sequence-to-sequence remaining useful life prediction method combining unsupervised LSTM encoding-decoding and temporal convolutional network. *Measurement Science and Technology*, *33*(8), 085013. https://doi.org/10.1088/1361-6501/AC632D

Khan, T., Ahmad, K., Khan, J., Khan, I., & Ahmad, N. (2022). An Explainable Regression Framework for Predicting Remaining Useful Life of Machines. *2022 27th International Conference on Automation and Computing: Smart Systems and Manufacturing, ICAC 2022*. https://doi.org/10.1109/ICAC55051.2022.9911162

Kobayashi, K., Almutairi, B., Sakib, M. N., Chakraborty, S., & Alam, S. B. (2023). *Explainable, Interpretable & Trustworthy AI for Intelligent Digital Twin: Case Study on Remaining Useful Life*. https://arxiv.org/abs/2301.06676v1

Kononov, E., Klyuev, A., & Tashkinov, M. (2023). Prediction of Technical State of Mechanical Systems Based on Interpretive Neural Network Model. *Sensors 2023, Vol. 23, Page 1892*, *23*(4), 1892. https://doi.org/10.3390/S23041892

Koutroulis, G., Mutlu, B., & Kern, R. (2022). Constructing robust health indicators from complex engineered systems via anticausal learning. *Engineering Applications of Artificial Intelligence*, *113*. https://doi.org/10.1016/j.engappai.2022.104926

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. https://doi.org/10.1145/2939672.2939874

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. https://doi.org/10.1016/J.YMSSP.2017.11.016

Li, A., Yang, X., Dong, H., Xie, Z., & Yang, C. (2018). Machine learning-based sensor data modeling methods for power transformer PHM. *Sensors (Switzerland)*, *18*(12). https://doi.org/10.3390/s18124430

Li, X., Ding, Q., & Sun, J. Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering and System Safety*, *172*, 1–11. https://doi.org/10.1016/j.ress.2017.11.021

Liu, B., Gao, Z., Lu, B., Dong, H., & An, Z. (2022). *SAL-CNN: Estimate the Remaining Useful Life of Bearings Using Time-frequency Information*. https://arxiv.org/abs/2204.05045v1

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. https://github.com/slundberg/shap

Maulana, F., Starr, A., & Ompusunggu, A. P. (2023). Explainable Data-Driven Method Combined with Bayesian Filtering for Remaining Useful Lifetime Prediction of Aircraft Engines Using NASA CMAPSS Datasets. *Machines*, *11*(2). https://doi.org/10.3390/machines11020163

Ramezani, S., Moini, A., & Riahi, M. (2019). Prognostics and Health Management in Machinery: A Review of Methodologies for RUL prediction and Roadmap. *International Journal of Industrial Engineering &*

Supply Chain Management, *6*(1), 38–61. www.ijiems.com

Remadna, I., Terrissa, L. S., Al Masry, Z., & Zerhouni, N. (2023). RUL Prediction Using a Fusion of Attention-Based Convolutional Variational AutoEncoder and Ensemble Learning Classifier. *IEEE Transactions on Reliability*, *72*(1), 106–124. https://doi.org/10.1109/TR.2022.3190639

Ren, C., Li, H., Zhang, Z., & Si, X. (2023). A Remaining Useful Life Prediction Method with Degradation Model Calibration. *Proceedings of 2023 IEEE 12th Data Driven Control and Learning Systems Conference, DDCLS 2023*, 172–177. https://doi.org/10.1109/DDCLS58216.2023.10166929

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85. https://doi.org/10.1214/21-SS133

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. *2008 International Conference on Prognostics and Health Management, PHM 2008*. https://doi.org/10.1109/PHM.2008.4711414

Serradilla Oscar, Ekhi Zugasti, Carlos Cernuda, Andoitz Aranburu, Julian Ramirez de Okariz, & Urko Zurutuza. (2020). Interpreting Remaining Useful Life Estimations Combining Explainable Artificial Intelligence and domain knowledge in industrial machinery. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8.

Si, X. S., Wang, W., Hu, C. H., & Zhou, D. H. (2011). Remaining useful life estimation – A review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14. https://doi.org/10.1016/J.EJOR.2010.11.018

Vollert, S., Atzmueller, M., & Theissler, A. (2021). Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, *2021-September*. https://doi.org/10.1109/ETFA45728.2021.9613467

Watson, D. (2020). Conceptual challenges for interpretable machine learning. *Synthese*, *200*, 1–33. https://doi.org/10.2139/SSRN.3668444

Xu, Q., Chen, Z., Wu, K., Wang, C., Wu, M., & Li, X. (2022). KDnet-RUL: A Knowledge Distillation Framework to Compress Deep Neural Networks for Machine Remaining Useful Life Prediction. *IEEE Transactions on Industrial Electronics*, *69*(2), 2022–2032. https://doi.org/10.1109/TIE.2021.3057030

Yan, B., Ma, X., Huang, G., & Zhao, Y. (2021). Two-stage physics-based Wiener process models for online RUL prediction in field vibration data. *Mechanical Systems and Signal Processing*, *152*. https://doi.org/10.1016/j.ymssp.2020.107378

Yang, F., Zhang, W., Tao, L., & Ma, J. (2020). Transfer learning strategies for deep learning-based PHM algorithms. *Applied Sciences (Switzerland)*, *10*(7). https://doi.org/10.3390/app10072361

Ye, Z., & Yu, J. (2023). A Selective Adversarial Adaptation Network for Remaining Useful Life Prediction of Machines Under Different Working Conditions. *IEEE Systems Journal*, *17*(1), 62–71. https://doi.org/10.1109/JSYST.2022.3183134

Zhang, K. ;, Liu, R., Zhang, K., & Liu, R. (2023). LSTM-Based Multi-Task Method for Remaining Useful Life Prediction under Corrupted Sensor Data. *Machines 2023, Vol. 11, Page 341*, *11*(3), 341. https://doi.org/10.3390/MACHINES11030341

Zhao, Y., & Addepalli, S. (2020). *ScienceDirect ScienceDirect Remaining Useful Life Prediction using Deep Learning Approaches: A Review*. https://doi.org/10.1016/j.promfg.2020.06.015

Zhou, J., Qin, Y., Chen, D., Liu, F., & Qian, Q. (2022). Remaining useful life prediction of bearings by a new reinforced memory GRU network. *Advanced Engineering Informatics*, *53*, 101682. https://doi.org/10.1016/J.AEI.2022.101682

Zhou, J., Qin, Y., Luo, J., & Zhu, T. (2023). Remaining Useful Life Prediction by Distribution Contact Ratio Health Indicator and Consolidated Memory GRU. *IEEE Transactions on Industrial Informatics*, *19*(7), 8472–8483. https://doi.org/10.1109/TII.2022.3218665

Zou, Y., Shi, K., Liu, Y., Zhangjidong, Liu, Y., & Ding, G. (2021). A Novel Machine RUL Prediction Method in Small Sample based on Fully Convolutional Variational Auto-Encoding Network. *2021 26th International Conference on Automation and Computing: System Intelligence through Automation and Computing, ICAC 2021*. https://doi.org/10.23919/ICAC50006.2021.9594066

BIOGRAPHIES

Yazan Alomari is a Ph.D. candidate in Computer Science at Eötvös Loránd University, Hungary, since September 2020, with collaboration from TU Delft University, The Netherlands. His research focuses on leveraging large language models for Prognostics and Health Management (PHM) in aerospace, emphasizing predictive maintenance through AI insights. His work spans across artificial intelligence, machine learning, and data mining & analysis, aiming to enhance system reliability and operational efficiency. Yazan has also contributed to academia as a University Lecturer at Eötvös Loránd University, imparting knowledge in data science, machine learning, and digital manufacturing. His publications underscore advancements in aerospace PHM, eXplainable AI and machine learning interpretation, showcasing his commitment to advancing AI applications in critical sectors.

Marcia L. Baptista BSc. and MSc. in Informatics and Computer Engineering. Instituto Superior Tecnico, Lisbon, Portugal September 2008) is an Assistant Professor at the Aerospace Faculty of TU Delft since 2020. She holds a PhD from the Engineering Design and Advanced Manufacturing (EDAM) program under the umbrella of MIT Portugal. Her research focuses on the development of prognostics techniques for aeronautics equipment. Her research interests include eXplainable Artificial Intelligence (xAI), machine learning, hybrid modeling, maintenance and prognostics.

Mátyás Andó habil., Ph.D., is Associate Professor in Institute of Computer Science, at Faculty of Informatics of Eötvös Loránd University, ELTE, Budapest, Hungary. He teaches: Manufacturing Technologies, CNC programing, CAM systems. His research area: Improve the efficiency in the industry, I4.0 solutions (CNC machines, PLC, Welding, Image processing). Material sciences (polymer development, tribology, 3D printing). Membership: Hungarian Association of Mechanical Engineers – GTE, members of the Hungarian Academy of Sciences (Section of Engineering Sciences).

Integration of Condition Information in UAV Swarm Management to increase System Availability in dynamic Environments

Lorenz Dingeldein[1]

[1] *Institute of Flight Systems and Automatic Control, Darmstadt, 64289, Germany*
dingeldein@fsr.tu-darmstadt.de

ABSTRACT

The approach of prognostics and health management (PHM) focuses on the real-time health assessment of a system under its actual operating condition and even extending this by the prediction of the future state based on up-to-date system information. This pursues the aim to derive more advanced maintenance or asset deployment strategies in order to keep the operation of the system safe and reliable. In this context, the outcome of a PHM system is often used as a decision support. For a high fidelity system where the actual state is considered at every timestep and a decision is executed immediately based up on this information, Reinforcement Learning (RL) becomes a tool to find an optimized solution. Therefore the paper presents a methodology that integrates health and operational data into a RL approach in order to derive immediate operational strategies for lower degradation and higher safety and reliability. The approach is evaluated on the basis of a swarm of unmanned aerial vehicles (UAVs) that performs a complete-area path-coverage (CAPC) mission. It can be shown that the integration of health information as well as environmental data describing dynamic operating conditions lead to lower degradation and result in more reliable operations of the swarm while achieving a more flexible mission performance compared to pre-divided swarm-missions. Varying states are also taken into account, which emphasises this approach to be a highly dynamic PHM system application.

1. INTRODUCTION

To avoid fatal incidents, safety and reliability are two major objectives for developments in the aviation industry (Tumer, 2011). While safety refers to system operations without causing harm or damage to people, property or the environment, reliability focuses on the ability of a system to perform its intended functions without failure or degradation over time (Stapelberg, 2009). The latter is the motivation to develop PHM functionalities where system states are predicted in order to derive decisions for maintenance strategies to increase reliability, availability and save costs. It is evident that usage leads to some level of degradation. Implementing a specific usage strategy can mitigate this degradation, resulting in extended system functionality and improved operational reliability, even in systems that have already experienced degradation. This approach is called prescriptive maintenance which complements PHM approaches by utilising their outputs, namely state detections and remaining useful life (RUL) predictions. While traditional PHM approaches try to extend system usage through a more precise calculation of the RUL, prescriptive maintenance guarantees for a reliable usage of systems that already show remarkable degradation (Marques & Giacotto, 2019). Combining usage specific degradation with a PHM based condition assessment is subject of this paper, which provides a prescriptive maintenance strategy using reinforcement learning.

The system used in this paper to implement the condition-based operational optimisation is a UAV swarm. The high-level mission goal of the swarm is the CAPC which applies to time-critical reconnaissance missions and also covers a common problem definition in the field of multi-agent (MA) robotics. The following reasons emphasize the suitability of this system and use-case for the developed approach:

- **Mission reliability:** System functionality of every swarm member needs to be guaranteed in order to be able to fulfill high level mission goals of a complete area coverage. While operational capabilities of a single UAV are limited, a swarm has the advantage of achieving more challenging mission goals in shorter time. The time-factor is crucial, for example, in search and rescue missions or forest fire observations.
- **Autonomy:** UAVs do not have a pilot on board. This means that various functionalities have to be automated. System state detection, as part of a PHM approach, is a decisive one in order to guarantee for reliable system functionality.
- **Redundancy:** Every individual swarm member represents a redundancy in the swarm structure. Individual

tasks can be distributed reasonable within the whole swarm in order to define a specific usage based upon environmental conditions. This guarantees a flexible task assignment that allows for optimization strategies.

In current research literature, the aspects of the integrated approach developed in this paper are considered separately. The basic idea of using PHM for a dynamic reliability assessment is described by the authors from (Heier, Mehringskötter, & Preusche, 2018). The paper emphasises the connection of PHM and reliability topics to develop decision support tools. The authors in (Bougacha & Varnier, 2020) use PHM as a driver for decision support. They pursue the primary goal of achieving higher reliability, availability and operational safety. Especially health and RUL indicators are utilized from the PHM approach in order to establish a decision-making process. Early approaches of in-mission decision making based on system states are discussed in (Andersson et al., 2015) and (Alighanbari, 2004). The latter even includes changes within the environment where the UAVs need to react to. In addition not only one UAV is part of the mission but a swarm of UAVs is considered. Data-driven approaches and machine learning techniques were not as easy accessible and developed as they are nowadays, leaving potential for the problems presented in these papers. A more recent consideration can be found in (Darrah, Quiñones-Grueiro, Biswas, & Kulkarni, 2021) where they use an online state observation to update parameters that optimize the prognosis for specific mission profiles. The better prognostic performance can than be used to derive more precise decisions but the focus of this paper is on a single UAV.

While the previously mentioned literature deals with linking PHM approaches with reliability, the following literature analysis focuses on deploying multi-agent swarm operations in a digitized environment. A baseline for multi-agent path-coverage is shown in (Cho, Park, Park, & Kim, 2021) where different grid-based map representations are compared. Even though hexagonal grids show certain advantages, such as increased navigation capabilities, the use of a cubic grid based map representation seems to be a suitable choice for the CAPC mission. Another approach for efficient swarm applications is described in (Mahmoud Zadeh, Yazdani, Elmi, Abbasi, & Ghanooni, 2022) and focuses on data acquisition. This approach could be interesting when deploying the swarm management approach from this paper in the real world and a good concept for data acquisition is needed. Nevertheless it describes the possibilities of UAV swarms. No CAPC is performed, but the distance between UAVs for better sensor measeruments is taken into account and considered as a useful approach. In (Radzki et al., 2021) travel uncertainties for a complete UAV fleet get determined. The result is used to optimize the usage of a UAV fleet but no in-mission decisions are made. This approach rather solves a scheduling problem than dealing with in-mission decisions to react on environmental conditions and system changes.

In order to make dynamic in-mission decisions, the approach of this paper uses reinforcement learning. This allows a large amount of heterogeneous data to be taken into account, which can change spontaneously in a sequential simulation. The successful application of reinforcement learning to a similar problem statement can be seen in the following literature. Using RL to control a swarm of buoys is described in (Kouzehgar, Meghjani, & Bouffanais, 2020). The goal is also the CAPC mission but input data differs in contrast to the deployment of UAVs. More comparable is the approach in (Puente-Castro, Rivero, Pazos, & Fernandez-Blanco, 2022) where a CAPC mission is performed with UAVs. The focus lies on the high level mission goal and enables the identification of relevant parameters for the coverage task. In addition (Xiao, Wang, Zhang, & Cheng, 2020) propose an approach to solve a CAPC task as well. No considerations of external factors or systems states are integrated into the approach but it helps to get an overview to solve the high level mission goal of CAPC. The closest approach is presented in (Theile, Bayerlein, Nai, Gesbert, & Caccamo, 2020) where power limitations of UAVs are integrated into a RL approach. The CAPC mission is specified as the target, but in fact more of a path-finding algorithm is implemented, which appears to be too permeable for reconnaissance missions and power limitations do not reflect the link between usage and degradation.

This paper uses the individual results from the literature stated above to develop an integrated solution for the condition-based organisation of a UAV swarm for the CAPC task using reinforcement learning. The general approach, including the experimental setup, is presented in Section 2. The results of the approach, applied to the described use case, are presented in Section 3 and analysed in Section 4. The paper concludes with a summary and an outlook on future work in section 5.

2. METHODOLOGY

To solve the task of a CAPC mission performed by multiple UAVs with respect to their health condition, a reinforcement learning method is implemented in python using RLlib (Liang et al., 2018). RLlib is only one of many possibilities to implement RL, whereby the following aspects are favourable for this paper:

- It is open-source
- It allows the integration of own simulation-environments
- It contains the option of implementing multi-agent reinforcement-learning (MARL) approaches

For the learning algorithm the Proximal Policy Optimization (PPO) is chosen. To date, this is the only MARL-capable algorithm in RLlib, so the implementation in comparable libraries should also be considered.

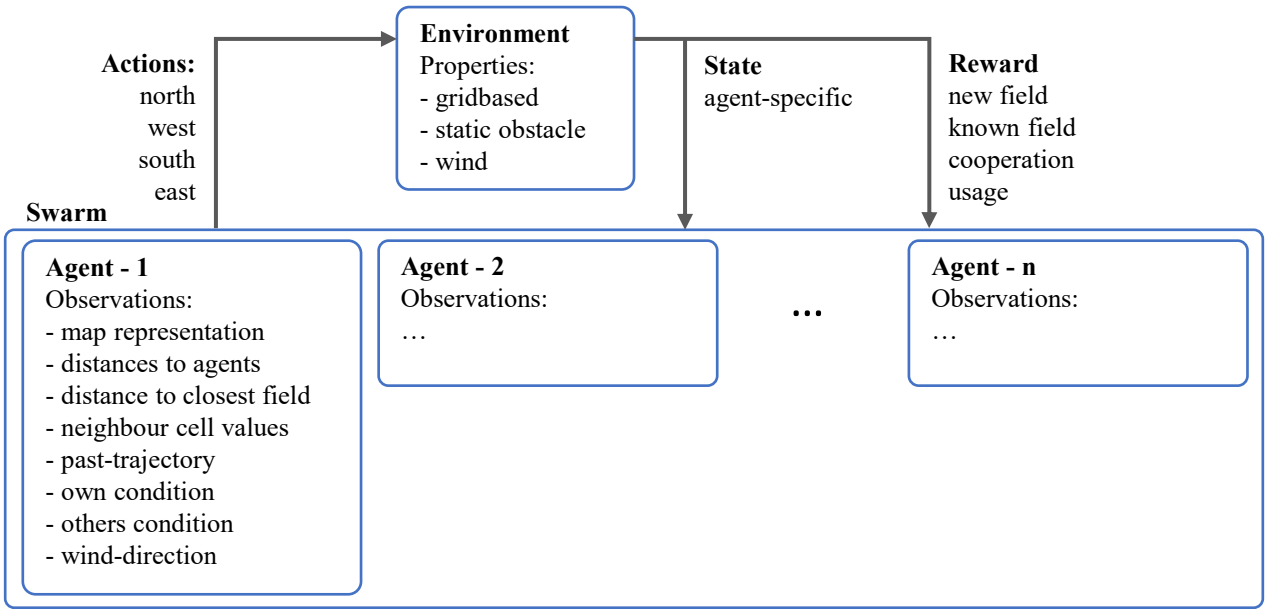


Figure 1. Framework for the multi-agent reinforcement-learning algorithm with consideration of system condition and environmental influences

The MARL method is similar to the standard RL framework where an agent is interacting with an environment through choosing actions and consequently receiving a reward. The basic principle of the MARL method with respect to the given task is shown in Figure 1. The methodology is based on the single-agent RL approach described in (Wiering & Van Otterlo, 2012) and extended with specifications to implement a MARL specific model. As it is the task to cover a certain area through creating a trajectory based on the movement decisions into the four main directions, a squared grid based environment is beneficial. This presupposes that a search is carried out along the search path with a certain radius, whereby simplified squares are assumed for coverage of a certain area that has been observed through a fly-over. While a third dimension could be used for deconfliction, it is neglected in this case to simplify the complexity of the system. The focus is on optimising a UAV swarm so that not only one agent interacts with the environment, but the actions of several UAVs are orchestrated and used as input for the environment. The action space thus becomes a vector that represents the four main directions of possible movements into north, east, south and west direction for every agent that takes part in the mission. The observation for every agent is derived from the state of the environment and takes agent-specific information into consideration. The reward rates the agents behavior and thereby helps the RL-algorithm to learn and successively improve the accumulative reward that is gathered in one mission. Every component of the MARL model is described in more detail in the following sections.

2.1. Environment Design

The MARL approach assumes multiple agents that cooperate together and interact with an environment through the execution of actions. The action of an individual agent changes the environment and in reverse calculates a reward that is fed back to the agent. In order to be aware about the actions to take the agent receives a state representation from the agent's point of view in form of observations. An environment model is necessary to provide a realistic and dynamic interaction between the environment and the agent, allowing the MARL model to learn and improve its decision-making through trial and error for a lot of training runs, which is ultimately understood as training process.

The use-case specified environment design is based on a grid based representation of a search field, that needs to be covered by a swarm of agents, the UAVs. The visualization of the environment used in this paper is shown in Figure 2 and subsequently described in detail.

The cell shape is square. This leads to four primary directions of movement, where the distance from the center of one cell to the center of its neighboring cells remains consistent. Cells that have been discoverd are displayed in beige, not visited cells are colored in green. Cells that are obstacles are colored as bricks and the wind direction is indicated as yellow arrows left and on top of the searchfield-cells. Assuming that the UAV proceed with a constant speed, the travelling distance of one timestep within the environment model is constant, which is also fits to the square shaped cells. When an UAV moves from one cell to another, it increments the value assigned to

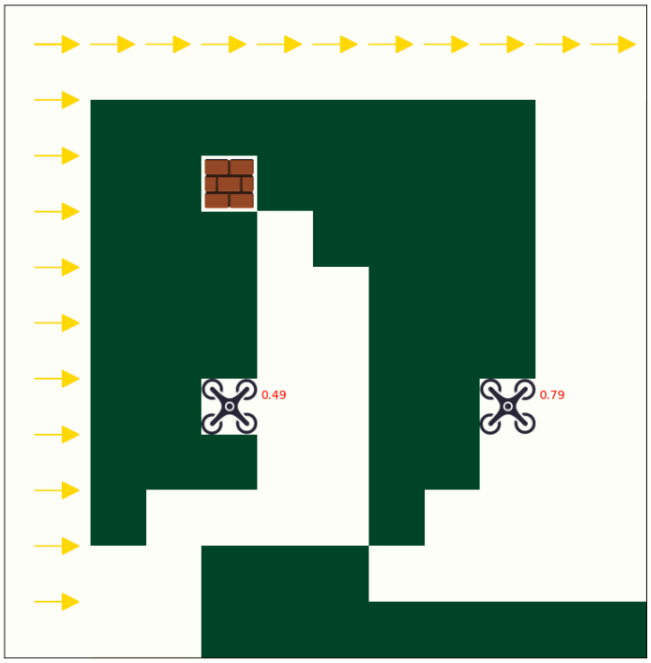


Figure 2. Environment design for a UAV-Swarm CAPC-Mission with external factors and varying system conditions

the cell by one, representing the number of visits to that cell. The highest value within the map is ten, which is utilized both for static obstacles and as a termination condition if an UAV visits a single cell too often. As an observation, which is described in detail in subsection 2.1.2, the UAV draws the map information and fortifies it to a potential map. In addition to the search field representation, the environment also incorporates a wind simulation. The wind simulation is used to specify the main influence on the system usage. A detailed description of the influence of wind on UAVs can be found in (Wang, Wang, Ali, Ting Ting, & Wang, 2019). It is assumed that the UAV is able to fly at constant speed under any wind condition. This results in a different power demand, depending on the direction of movement of the UAV and the prevailing wind direction. Higher power consumption means greater stress on the components and therefore increased degradation. The swarm configuration enables a system management where degraded UAVs take over the coverage of the search field crosswind and are thus exposed to lower degradation, while intact UAVs can take over more demanding trajectories against the wind and can absorb higher degradation without noticeable increasing the risk. The risk mitigation can be derived from the general assumption of degradation taken from (Kim, An, & Choi, 2017). While at the beginning of a system life the degradation is mainly characterised through wear represented through a linear progression, the degradation grows exponentially at the end of a system life, which results in higher chances for unexpected system failures.

Within the environment design wind is considered as constant while the direction of wind can change between the missions. Generally missions are treated individually so that the condition of one UAV gets defined at the beginning of a mission as well. The condition is chosen arbitrarily between 0.1 and 1, representing UAVs with high degradation when a low value is chosen and UAVs with a good condition if a value is chosen that is close to one.

The upcoming sections explain the design of the remaining parts of the MARL approach: the action space, the observation space, and the reward function. Additionally, it covers how missions are initialized and outlines the experiments conducted to assess RL-algorithm performance.

2.1.1. Action Space

The grid-based representation of the search field enables movements along the four main directions, from cell center to cell center, ensuring an equal traveling distance. The actions are defined as a single value from zero to three, representing the four main directions. The movement than is executed on a global scale, which means that based on the chosen value from the movement vector, the UAV moves to the north, west, south or east. In contrast, an UAV-centred approach could be chosen, which changes the direction of flight depending on the chosen action previously. In this case, four actions would also be conceivable, one value for continuing to fly straight ahead, one value each for a left or right turn and one value for reversing the direction of flight. In the remainder of the paper, however, the global approach is pursued further.

2.1.2. Observation Space

The UAVs draw an self-centered observation from the environment after changing it with their action. The observation space contains the following:

- **Map representation**: The UAV gets a matrix that counts the visits of the fields. In addition other UAVs as well as obstacles are highlighted with a value of the maximum allowed visits for one run. The latter is used for the termination condition and is described in more detail in section 2.1.4.
- **Own position**: The UAV gets its current position after it moved. Because of the two dimensional characteristics of the environment the position is represented globally as a xy-position within the environment grid.
- **Distances to UAVs**: The UAV gets the distances in number of cells in xy-direction to all other UAVs that are operating for one mission.
- **Field distances**: The UAVs gets the distances in number of cells in xy-direction to the closest field that does not count any visit from any UAV.
- **Surrounding**: With the surrounding data the UAVs get a representation of the environment based upon their cur-

rent position. With a size of four by one the surrounding matrix contains the values of the local map representation based on the UAVs position into the four main direction. If the UAV operates close to the border of the search field, values that exceed the search field are represented with a value of ten which is equal to the value of obstacles or other UAVs which should not be visited and lead to a mission termination.

- **Movement history**: The movement history is a vector of the length of five which contains the direction decisions of the UAV in a chronological sequence. For every step of the UAV within the environment the last value of the sequence is deleted, the sequences and a the direction that the UAV moved during this step is added as the first entry to the vector.
- **Own condition**: The UAV needs to know its own condition to compare it with the condition of the other UAVs. It is a normalized value between zero and one and saved as a scalar in the observation space.
- **Others condition**: Due to the same reason as before the observation space of a single UAV also contains all the conditions of the other UAVs that are participating in the swarm mission.
- **Wind information**: As the wind information mainly is responsible for the usage and degradation of the UAV it is also integrated into the observation space as a two dimensional, directional vector.

Based on the observations of the environment and the UAVs behavior the RL-algorithm is able to coordinate the UAVs movements with respect to dynamic environmental states. The goal is to reduce intensive usage for UAVs with bad condition, which gets then compensated by the UAVs that are in good condition. This results in an overall lower degradation according to (Kim et al., 2017) where it is stated that the degradation of a system increases over usage time in two steps, firstly linear and afterwards exponentially. Further more it reduces the risk of sudden system failures which occur with a higher chance to the end of life of a system and therefore decreases mission risk and increases mission reliability. To allow the UAV to optimize its decisions it receives a reward after every step according to section 2.1.3. The UAVs also exchange and communicate information about their position and condition, enhancing their decisions even further, establishing swarm intelligence.

2.1.3. Reward Function

In order to get the UAV to perform as desired, it receives a reward based upon its decision and the changes that occur within the environment at every step and at the end of the mission. The major goal in the mentioned use-case is to completely cover a designated area. The sub task consists of the efficient coverage of this area with respect to the systems usage that in combination with the degradation state has an impact on mission reliability. Therefore the reward can be divided into two types. The step-wise reward that is applied at every step on every UAV and a sparse reward (Hare, 2019) that is applied at the end of a mission.

The step-wise reward consists out of a positive reward for visiting unseen cells of the environment. For every new cell the UAV receives a reward of $R_{new\_visit} = 1$. That causes the reward to increase linear for the exploration of new fields. This motivates the UAV to search for isolated cells. The UAV receives a reward of $R_{already\_visited} = -1$ if the cell was already visited. This encourages to discover new cells as fast as possible.

To support cooperative search the UAVs get an additional positive reward R_{coop} every step if there is more than one UAV active to complete the mission. This means that the swarm has to organize itself based on the environment representation and without crashing into an obstacle, another UAV or the boundary of the searchfield in order to achieve this reward. The formula for R_{coop} is as follows:

$$R_{coop} = \begin{cases} 1 & n_{UAVs\_active} > 1 \\ 0 & else \end{cases} \tag{1}$$

The reward function incorporates both the UAVs' conditions (AC) and usage conditions by comparing the direction of movement with the wind direction. It is assumed that flying against the wind (headwind) is more energy consuming than flying crosswind or with the wind (tailwind). Therefore an angular comparison of wind direction and movement direction is made and energy cost (EC) for the manoeuvre gets calculated as following:

$$EC = \begin{cases} 0 & Tailwind \\ 0.5 & Crosswind \\ 1 & Headwind \end{cases} \tag{2}$$

It is only possible to use this energy cost if a constant speed is assumed. This is also helpful for integrating the system behaviour into a reinforcement learning environment. This simplification is made in order to focus on and analyse the interactions of degraded systems and environmental conditions within the multi UAV environment. To link environmental conditions and degradation, the reward is conditionally calculated as follows:

$$R_{usage} = \begin{cases} 1 & AC > 0.5 \; and \; EC = [0,1] \\ -5 & AC > 0.5 \; and \; EC = 0.5 \\ -1 & AC < 0.5 \; and \; EC = [0,1] \\ 5 & AC < 0.5 \; and \; EC = 0.5 \end{cases} \quad (3)$$

It shows that R_{usage} is dependent on the wind direction and the UAV condition. An UAV in good condition is meant to fly against the wind. This means that the UAV has to fly with the wind after a certain amount of steps in wind direction in order to not leave the search field (which leads to a mission termination). For this reason, it is not possible to differentiate between flying with and against the wind. Accordingly the flight movements of the UAV in bad condition must inevitably take place increasingly in crosswinds to reduce the proportion of movement directions against the wind.

The cumulative reward values calculated for each UAV at every step are aggregated over an episode which stands for a mission until a termination criteria is met. Initially, rewards are determined individually for each UAV, and at the episode's end, the total rewards across all UAVs are summed up. The end of an episode is initiated by predefined termination conditions. An additional reward known as the sparse reward is introduced alongside the termination condition, both of which will be further detailed in the following subsection.

2.1.4. Termination Conditions

Termination conditions are necessary to end an episode which is equivalent to a mission. They can be triggered if the mission task is fulfilled, the UAV's behavior leaves specified boundaries or to prevent inefficiency where the episode is trapped in an infinite loop. With the problem at hand, the termination conditions are chosen as follows:

- **Completely covered:** The UAVs were able to visit every cell of the designated search field at least once. For completing the task the UAVs do not get a negative reward. This can also be interpreted as a sparse reward that motivates the UAV to perform the task as efficient as possible with regard to the coverage performance. If an UAV is not active at the end of an episode, it gets a negative reward as described in the crash termination condition.
- **Inefficient search:** The episode gets cancelled if one of the cells within the search field gets visited more than ten times. In that case the sparse reward is -100 minus the number of unexplored fields of the search field. This reward applies for every UAV of the swarm that is still active at that time. Otherwise the crash termination. Otherwise, crash termination has already been applied.
- **Crashes:** It is classified as a crash if an UAV shares the same cell with another UAV, an obstacle or if it leaves the search field. In that case the sparse reward is calculated the same way as it is calculated for the inefficient search and the crashed UAV stops exploring the search field.

The primary objective of the MARL approach is to maximize the accumulation of rewards within a single episode such that the reward function significantly determines the behaviour of the UAVs. Within section 4 the effects of changing the reward function will be discussed in detail.

2.1.5. Initialization

Certain initial conditions must be defined to start the simulation. This includes:

- **Number of UAVs:** The primary focus of the RL-algorithm pertains to the optimization of the concurrent operation of multiple UAVs. The parameter dictating the swarm size can be specified during the initialization phase.
- **Starting location of UAV:** The UAVs are meant to fly to the designated search field, therefore their starting position is always at the boarder of the search field. To maintain a certain distance to each other, every UAV starts from another side of the search field, representing different UAV bases and approaching directions.
- **UAV condition:** The condition of the UAV is determined through a random selection process within the interval of 0.1 to 1, with precision of two decimal places.
- **Map representation:** A map in form of an array, representing the search field coverage, that counts the visits of each cell. The map is adjusted during the course of the mission as described in section 2.1.
- **Wind direction:** An initial wind direction is defined in form of a two dimensional vector.
- **Obstacle position:** While the UAVs can be understood as moving obstacles, fixed obstacles are also defined within the initialization phase as high values in the map representation.

The parameters during initialization are adaptable to specific requirements, facilitating the experimentation and evaluation of the RL-algorithm across diverse scenarios. The next section provides detailed explanations on how the parameters are set up for the experiments conducted in this paper.

2.2. Design of Experiments

A Monte Carlo simulation was run to assess the capability of reinforcement learning in optimizing specific relationships, particularly focusing on the dynamic management of UAV degradation in response to varying environmental conditions during the execution of a CAPC mission. The experiments are set up almost with the same parameters but are randomized with the regard to the following parameters:

- UAV starting position

- Obstacle location
- Wind direction
- UAV condition

The training parameters such as number of episodes, batch size and RL-algorithm are chosen as it is proposed by the documentation of the Python library of RLlib. For the evaluation, the trained RL-algorithm that achieved the best result is used, which can be determined by analysing the average reward of the learning curve (3.1). The experiment consists out of 100 runs. The metrics used for the evaluation of the experiment is described in the following section.

2.3. Evaluation metrics

Two metrics are used for the evaluation of the experiments. The first metric counts the cells with an equal number of visits using the following pythonic algorithm:

Algorithm 1 Evaluation of Coverage Performance

```
cell visits = [0 for visits in range(0, max(visits))]
for cell in searchfield do
    if cell is not obstacle then:
        cell visits[cell in searchfield(visits)] += 1
```

The list of cell visits is then visually represented and should give evidence about the coverage performance of the trained RL-algorithm. The visualization can be seen in Figure 4 for the coverage performance of a completely trained RL-algorithm where the results for 100 missions are summarized with the help of errorbars. The goal is to avoid multiple visits of cells which shortens the mission time for complete area coverage.

The second metric compares the movement decisions made by the trained RL-algorithm based on the UAVs condition. The evaluation is performed using the following formula:

$$UAV\ Wind\ Load = \begin{cases} Headwind & WD \angle MD = 180° \\ Crosswind & WD \angle MD = \pm 90° \\ Tailwind & WD \angle MD = 0° \end{cases} \tag{4}$$

The case differentiation of loads the UAV experiences based on the wind is determined by calculating the angle between the wind direction (WD) and the direction of movement (MD). The UAV wind load can be linked to the UAV state and can thus be visualised in a bar chart (see Figure 5). By comparing the frequency of movement decisions in connection with the UAV state, it is possible to assess whether the RL learner has learnt to use the UAV swarm as efficiently as possible with a focus on the system-state.

3. Results

The following section presents the results from the experiment described in 2.2. First, the overall training process is pictured. This is followed by the results of the coverage performance, which enable the evaluation of the first sub-task of the RL approach. The results of the second sub-task are presented in the last subsection, showing an evaluation that focuses on the cooperation and degradation of the UAVs within the RL approach.

3.1. Learning performance

The goal of the reinforcement learning process is to increase the average reward successively over the number of training iterations. While a supervised learning approach compares the produced output of a network with labeled data and backpropagates the error, RL does not need labeled data and it is producing training data within the learning process. The reward function helps to choose the actions that lead to the best reward. This is not only considered at every single step within one training episode, but also at the end of one episode to increase the cumulative reward. The reward function used in this paper (described in subsection 2.1.3) should establish a multi-UAV cooperation to perform a complete area path coverage with respect to degradation that results from the individual system usage. The average reward achieved by the RL approach over the training process can be seen in Figure 3.

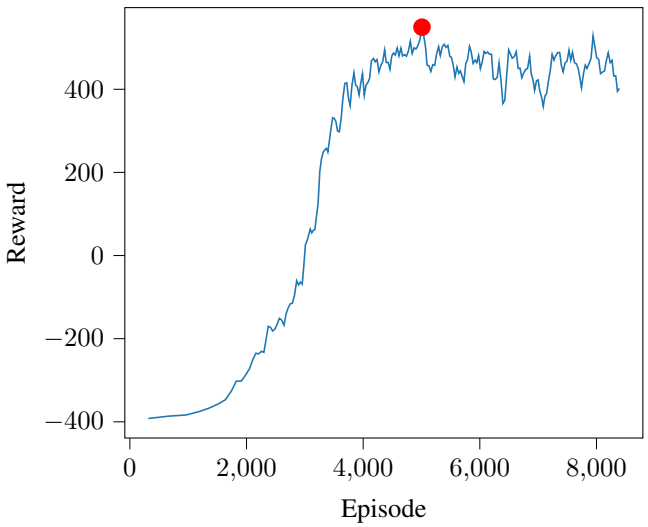


Figure 3. Cumulative training reward achieved by the RL approach over the training process

The training process starts with a high negative reward, which is comprehendable because the movement of the UAV is arbitrary and due to the starting position at the boarder of the search field, the UAV leaves the search field quite often at the very beginning of an episode, resulting in a high penalty and

low gathered reward for covering new fields and moving with usage considerations. The UAV then learns to not directly leave the search field but rather fly in a straight line until it reaches the opposite boarder. The average reward per mission increases slightly, which can be seen at the beginning of the reward curve. Subsequently the UAV learns to move in the right wind direction, paying respect to its own condition. It also learns to change direction at the boarder of the search field, resulting in a much higher average reward. This can be seen from the exponential increase in the reward curve. Afterwards it is harder for the UAV to consider the movement of the other UAVs, still it is able to optimize its movement pattern with respect to wind, information about the rest of the swarm and surrounding map data. The increase in the average rewards achieved per episode decreases again, whereby the reward curve reaches a saturation point. The convergence behavior at the end of the training does show instability, which can be explained by varying coverage and cooperation performance. Nevertheless, it can be concluded from the amount of the reward at the end of the learning process and the consideration of the reward function that the UAVs can achieve the first sub-goal in co-operation, namely searching the search field with slightly varying performance. Using the metrics that are described in 2.3 the coverage performance is discussed in more detail in the next section, as well as the level of cooperation where the developed metrics give more insight about the RL-algorithm performance.

3.2. Coverage performance

The primary goal is the CAPC. Only if the UAV is able to fulfill this kind of mission the cooperation performance with respect to the swarm condition can be compared and evaluated. To evaluate the coverage performance not only the complete coverage is considered but also the effectiveness of the coverage through counting the number of visits per cell. However, because this ideal solution conflicts with a search that takes environmental and systems conditions into account, coverage performance varies slightly at the end of the mission and cells of the search-field are visited more than once. Nevertheless the MARL approach is able to complete cover the search field area 92% of the time. This is not ideal but enough to evaluate the RL-algorithm performance. The result of the cell visit counts in order to evaluate the coverage performance can be seen in Figure 4.

The figure shows the number of cell visits on the x-axis and the frequency of occurrence of cells with the number of visits (from the x-axis) for a completed mission on the y-axis. The RL-algorithm was completely trained according to Figure 3. To get a representative behavior of the trained RL-algorithm 100 missions were performed for evaluation. The display with error bars clearly shows that the coverage performance is in a very good range. This is illustrated by the very low number of missions in which fields with zero visits

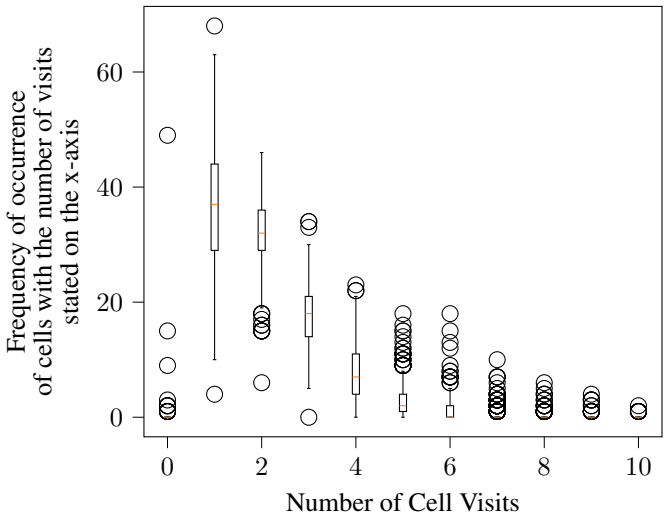


Figure 4. Evaluation of coverage performance by counting the number of fields with the same number of visits

remain at the end of a mission. The fields with zero visits are also categorised exclusively as outliers. A lot of outliers are also visible at fields with a high number of visits, which is also beneficial, because it means that the UAV learns that it should visit a single field as little as possible. This statement is confirmed by the highest value for single field visits. Overall, the distribution of field visits takes the form of a Weibull distribution that is used to describe the frequency of wind speed. Weibull distributions are also often used to describe the lifetime of technical components. Both aspects, namely wind and system lifetime are present within the presented framework and it is remarkable to see that the trained UAV shows such a behavior. Further analyses of the relationship between UAV behaviour and the Weibull distribution are pending.

3.3. Cooperation and degradation evaluation

The secondary goal of the trained RL approach is to coordinate multiple UAVs such that they are utilized according to their condition. This should encourage a usage suitable deployment of the swarm members in order to avoid sudden system breakdowns and increase reliability for the whole mission. To evaluate system usage with respect to environmental conditions, the number of movement decisions depending on the wind direction where the UAV conditions differ at least about 0.5 is counted. The result can be seen in Figure 5.

The barchart shows the decision of the UAVs with bad condition in blue and the decisions of the UAVs with good condition in orange. Only the values for which both UAVs were active are used, as otherwise cooperation is not possible. Furthermore, the values are normalised so that they can be easily compared with each other. It can be seen that the weaker UAV

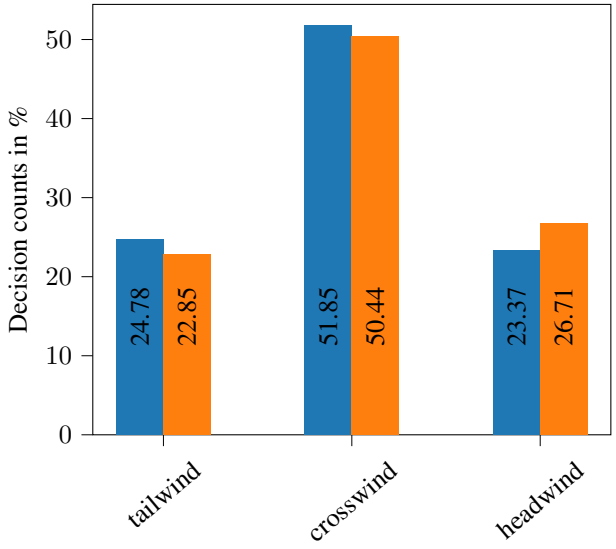


Figure 5. UAV decisions within a cooperative environment

chooses to move more into the direction in which it experiences crosswind. This part equally escapes the movement decisions against and with the wind. On one hand this suits the reward function. On the other hand this leads to less degradation for the weak UAV where it avoids moving against the wind. Choosing the major moving direction crosswind also avoids the UAV to fly against the wind after travelling a long distance with tailwind. Defining the right reward function is sometimes contradicting which gets discussed in the next section.

4. Discussion

During the implementation of the MARL approach, challenges arose with regard to the reward design and the superimposed objectives within the MA mission, which will be discussed below.

4.1. Reward Design

The reward design is very sensitive to minor changes. Also the weighting of the reward significantly changes the behavior of the UAVs. Not all intuitive rules for the reward achieve the desired effect, as the RL-algorithm incorporates the numerical values directly into its learning process. This is also partly dependent on the environment design. As shown in Section 3.3, no reduced degradation can be achieved by flying with the wind, as this inevitably requires flying against the wind from the search field boundary onwards. Another example is a weighted negative reward for multiple cell visits. It could be assumed that if not only a constant negative reward is used for cells visited several times, but the negative reward is multiplied by a factor derived from the number of visits per cell, better UAV behaviour is achieved. However, this is not the case, as the UAV is restricted in its free movement across the search field. Reaching an unvisited field directly would be associated with an increased negative reward. In order to find the right reward policy, the paper used a trial and error approach, so that there is further potential for optimisation at this point. This can also be realised through a different environment design that is connected with the reward assignment.

4.2. Superimposed Objectives

The MARL approach in this paper combines two goals, which creates a conflict between objectives. Both goals can only be achieved if compromises are made with regard to the individual goals. On the one hand, this complicates the reward design that comes into play at the end of a step. On the other hand, it makes evaluation methods more difficult. As this paper is a proof of concept and the assessment of performance is not the main focus, the topic of detailed evaluation should be the subject of further work.

5. Conclusion

This paper presents a MARL approach to solve a CAPC mission under the consideration of dynamic system states and other external factors which places a stress on the deployed systems. The topic dealt with is motivated by the reference to current research topics and specified by analysing the relevant research literature. A generalised methodology is derived that allows state and environment data to be integrated into a MARL approach. This approach allows individual UAVs to communicate with each other and perceive their surroundings as they navigate through the environment. The emphasis lies on designing the reward function, as it serves as the primary driver influencing the behavior of the UAVs, which is intended to utilize the swarm members in a resource-saving manner as an approach for optimisation.

A drone reconnaissance mission is used as a practical example to apply all components of the generalized methodology. The RL-algorithms performance is then evaluated regarding the learning process and the RL-algorithm performance. It can be stated that the completely trained RL-algorithm is able to solve the superimposed objectives of covering the complete area under consideration of the varying system state of the UAVs and a varying wind direction as an external factor. Through the integration of system condition and external loads through wind, the system usage is the main parameter that gets optimized. It turns out that due to the conflicting goals and the associated reward function, the behaviour of the UAVs follows a compromise. While the coverage performance decreases slightly, a more energy-efficient use of the drone swarm can be observed. With that, the methodology is able to recover from sudden system failures and guarantee a more reliable mission fulfilment. This extends existing approaches from current research literature through a highly dynamic in-mission decision process. In addition, a much freer

mission design is made possible by dispensing with segmentation of the search field.

With the promising results of this paper, an in-depth analysis of RL-algorithm performance based on relevant parameters is pending. Such an analysis can provide further insight about the design of the reward function and thus help to design the MA system for a desired behaviour.

NOMENCLATURE

AC	agent condition
$CAPC$	complete-area path-coverage
EC	energy consumption
MA	multi-agent
$MARL$	multi-agent reinforcement-learning
PHM	prognostics and health management
PPO	Proximal Policy Optimization
RL	reinforcment learning
RUL	remaining useful lifetime
UAV	unmanned aerial vehicle

REFERENCES

Alighanbari, M. (2004). *Task assignment algorithms for teams of uavs in dynamic environments* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Andersson, K., Bang, M., Marcus, C., Persson, B., Sturesson, P., Jensen, E., & Hult, G. (2015). Military utility: A proposed concept to support decision-making. *Technology in society*, *43*, 23–32.

Bougacha, O., & Varnier, C. (2020). Enhancing decisions in prognostics and health management framework. *International Journal of prognostics and health management*, *11*(1).

Cho, S.-W., Park, J.-H., Park, H.-J., & Kim, S. (2021). Multi-uav coverage path planning based on hexagonal grid decomposition in maritime search and rescue. *Mathematics*, *10*(1), 83.

Darrah, T., Quiñones-Grueiro, M., Biswas, G., & Kulkarni, C. S. (2021). Prognostics based decision making for safe and optimal uav operations. In *Aiaa scitech 2021 forum* (p. 0394).

Hare, J. (2019). Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*.

Heier, H., Mehringskötter, S., & Preusche, C. (2018). The use of phm for a dynamic reliability assessment. In *2018 ieee aerospace conference* (pp. 1–10).

Kim, N.-H., An, D., & Choi, J.-H. (2017). Prognostics and health management of engineering systems. *Switzerland: Springer International Publishing*.

Kouzehgar, M., Meghjani, M., & Bouffanais, R. (2020). Multi-agent reinforcement learning for dynamic ocean monitoring by a swarm of buoys. In *Global oceans 2020: Singapore–us gulf coast* (pp. 1–8).

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., … Stoica, I. (2018). Rllib: Abstractions for distributed reinforcement learning. In *International conference on machine learning* (pp. 3053–3062).

Mahmoud Zadeh, S., Yazdani, A., Elmi, A., Abbasi, A., & Ghanooni, P. (2022). Exploiting a fleet of uavs for monitoring and data acquisition of a distributed sensor network. *Neural Computing and Applications*, 1–14.

Marques, H., & Giacotto, A. (2019). Prescriptive maintenance: Building alternative plans for smart operations. In *The 10th aerospace technology congress.*

Puente-Castro, A., Rivero, D., Pazos, A., & Fernandez-Blanco, E. (2022). Uav swarm path planning with reinforcement learning for field prospecting. *Applied Intelligence*, *52*(12), 14101–14118.

Radzki, G., Bocewicz, G., Golińska-Dawson, P., Jasiulewicz-Kaczmarek, M., Witczak, M., & Banaszak, Z. (2021). Periodic planning of uavs' fleet mission with the uncertainty of travel parameters. In *2021 ieee international conference on fuzzy systems (fuzz-ieee)* (pp. 1–8).

Stapelberg, R. F. (2009). *Availability and maintainability in engineering design*. Springer.

Theile, M., Bayerlein, H., Nai, R., Gesbert, D., & Caccamo, M. (2020). Uav coverage path planning under varying power constraints using deep reinforcement learning. In *2020 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 1444–1449).

Tumer, I. (2011). System health management: With aerospace applications. In *Chapter* (Vol. 8, pp. 129–142). John Wiley and Sons United Kingdom.

Wang, B. H., Wang, D. B., Ali, Z. A., Ting Ting, B., & Wang, H. (2019). An overview of various kinds of wind effects on unmanned aerial vehicle. *Measurement and Control*, *52*(7-8), 731–739.

Wiering, M. A., & Van Otterlo, M. (2012). Reinforcement learning. *Adaptation, learning, and optimization*, *12*(3), 729.

Xiao, J., Wang, G., Zhang, Y., & Cheng, L. (2020). A distributed multi-agent dynamic area coverage algorithm based on reinforcement learning. *IEEE Access*, *8*, 33511–33521.

BIOGRAPHY

Lorenz Dingeldein received his M.Sc. in mechanical and process engineering from Technische Universität Darmstadt, Germany, in 2019. Since then he has been working at the Institute of Flight Systems and Automatic Control (FSR). As a Research Associate, he has been involved in several projects with a strong focus on Prognostics and Health Management. He focuses on the development of condition based asset management of multi agent systems in dynamic environments.

Labeling Algorithm for Outer-Race Faults in Bearings Based on Load Signal

Tal Bublil[1, *], Cees Taal[2], Bert Maljaars[2], Renata Klein[3], Jacob Bortman[1]

[1] *PHM Laboratory, Department of Mechanical Engineering, Ben-Gurion University of the Negev, P.O.B653, Beer-Sheva 8410501, Israel*

talbub@post.bgu.ac.il
jacbort@post.bgu.ac.il

[2] *SKF, Research and Technology Development, Meidoornkade 14, 3992AE, Houten, the Netherlands.*

cees.taal@skf.com
bert.maljaars@skf.com

[3]*R.K. Diagnostics, P.O. Box 101, Gilon, D.N. Misgav 20103, Israel*

Renata.Klein@rkdiagnostics.co.il

ABSTRACT

Rolling element bearings are essential components for the proper functioning of many types of rotating equipment. Diagnosing faults in bearings has traditionally been done using signal processing techniques inspired by physics, wherein acceleration signals are analyzed using time-frequency analysis methods. To study the effect of bearing damage on acceleration signals, experiments are typically performed aiming for a natural propagation of a spall. However, the extent of spall severity during the test remains uncertain. It is possible to disassemble and reassemble the bearing for visual inspection. Nevertheless, previous studies observed that the vibration signal would drastically change if this operation was conducted repeatedly, impacting the identification of trends in the acceleration signal. The objective of this study is to provide a method which can assist with labeling the spall size in endurance tests without the necessity of disassembling and reassembling the test rig. To address this issue, a new algorithm, based on the load cell signal was developed to assess the spall size using low-speed measurements. This algorithm enables the identification of the circumferential angle at which the rolling element interacts with the spall and is only carrying a partial load. The algorithm has been validated through visual inspections conducted during the experiment. This algorithm makes it possible to estimate the spall size without the need for visual inspection in subsequent experiments. A labeled endurance test contributes to a better understanding of spall propagation, such as the effect of speed, load, and material properties on the propagation speed. This study demonstrates how the load signal can be used for fault labeling with relatively simple and common techniques. This approach will enable the tackling of advanced and more complex problems in future endeavors, such as fault severity estimation and even prognosis.

1. INTRODUCTION

Bearings play a crucial role in nearly all rotating machinery (Malla & Panigrahi, 2019), and monitoring their condition typically involves four stages: detection, identification, severity estimation, and prognosis (Bechhoefer & Schlanbusch, 2018). A substantial amount of research has been conducted on the subject, with a focus on the detection and identification stages, which have shown promising results. Bearing damage severity is typically defined as a function of overall vibration levels ((ISO), 2016). Unfortunately, these thresholds are very application dependent and difficult to generalize. A robust and objective method for severity estimation in bearings remains challenging.

One common approach to define severity is based on raceway spall size in the circumferential direction. Among the studies engaging spall size estimation, various types of sensor measurements are utilized. Common methods include accelerometers, some employing oil debris monitoring (ODM), and others utilizing optic fibers (Gazizulin et al., 2019; Madar et al., 2022; Medvedovsky et al., 2022). A review of different approaches is given in Zhang et al., 2022. Accelerometers are relatively common components in

machinery due to their ease of installation. Methods based on accelerometers typically try to identify the entry and exit points of rolling elements from the spall within the time domain (Epps I K, 1991; Moazen-ahmadi & Howard, 2016; Sawalhi & Randall, 2011). Some of these studies employ low-pass filters to detect entry and exit events (Moazen Ahmadi et al., 2016), a practice that may pose challenges due to its reliance on a rule of thumb. Additionally, these methods are often detecting very small defects, whereas numerous applications involve significant spall sizes. There are also severity estimation methods which utilize condition indicators, which are then employed to calculate a health indicator (Gebraeel et al., 2004; Ma et al., 2012). These methods offer increased robustness in noisy conditions compared to the aforementioned methods since they consider trends based on multiple sensor recordings over time. However, establishing a direct link between health indicators and spall size in rotating machinery poses challenges, primarily due to missing ground truth values of its spall size during operation.

Studies that use ODM can estimate the spall size by calculating the total mass of debris particles originating from the bearing (Madar et al., 2022; Portal et al., 2022). However, to use this method, certain geometric assumptions are made which might be invalid. Optic fibers are used to measure the strain on the housing bearing, and by tracking the changes in the signal, it is possible to calculate the length of the spall (Medvedovsky et al., 2022). Nevertheless, both of these methods require expensive equipment and are not suitable for every test rig or machinery.

Emulating the topography of a spall is a challenging task. Consequently, studies that have explored severity estimation in bearings often rely on artificial spalls with less realistic rectangular shapes. However, the interaction between the rolling element (RE) in the bearing and the artificial spall could be significantly different from the interaction with a real spall, which may result in higher impulses in the acceleration signal than those observed in natural spalls (Zhang et al., 2021).

One approach to achieve naturally growing spalls is through endurance tests. Nevertheless, for measuring the spall size in acceleration algorithm validation, it is necessary to disassemble and reassemble the test rig, which can significantly alter the vibration signal (Smith & Randall, 2015). Recent studies have shown, that for specific test rig setups a load cell can act as a proximity measurement for displacement containing a distinctive pattern related to the spall geometry (Zhang et al., 2022). Moreover, an observation is made that the load-cell signal is less sensitive to the re-assembly of a bearing compared to acceleration.

In this study we propose a unique algorithm to estimate spall sizes in endurance tests using load signals, which can be obtained without visually inspecting the spall. The algorithm is implemented at low speeds, enabling validation of spall dimensions during endurance experiments.

2. Experimental Setup

The endurance test was conducted in SKF Research and Technology Development (RTD). The test was performed on the R2 test rig (Harris, 2006), as shown in Figure 1, with the positions of the accelerometer and the load cell indicated. For measuring the rotational speed, a tachometer measuring two pulses per shaft rotation was used. In the experiment, two bearings were measured: the tested bearing, located on the left side of the test rig, and a reference intact bearing positioned on the right side of the test rig. Both bearings were monitored throughout the experiment. The algorithm developed in this study was primarily validated using the load-cell data acquired from the tested bearing on the left side. Throughout the experiment, sensor snapshots were recorded, defined as synchronized recordings of all sensors at a sample rate of 49152 Hz for 36 seconds.

To validate the load-based algorithm, a visual inspection was required. To simplify the process of disassembling and reassembling the test rig, a bearing with a design that allows easy access to the outer race was chosen. The selected bearing is a cylindrical roller bearing of the N209 ECP type. During the experiment, only a pure radial load is applied because this bearing type cannot sustain axial loads.

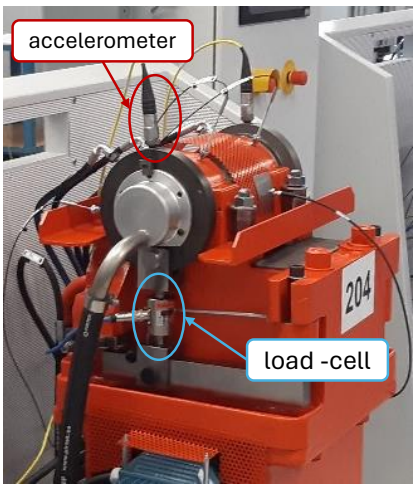


Figure 1: SKF R2 test rig; The accelerometer marked in red, load cell marked in blue.

2.1. Test Procedures

The test consists of two stages: (1) a damage initiation phase and (2) a spall growth phase. Both phases will be further explained in the following sections.

2.1.1. Damage Initiation Phase

In this phase, the purpose is to initiate a spall on the outer race of the bearing. To expedite this process, an initial small

damage was introduced to the outer race before the test. The damage was created using an electrical discharge machine (EDM) on the race surface. The EDM created a rectangular-shaped damage with circumferential and axial dimensions of 0.2mm and 2mm, respectively. In this phase, the bearing was subjected to high loads and speeds to induce growth. The decision to stop this phase was made based on an acceleration-based condition indicator (Harris, 2006), where an abnormal increase determined the stopping criteria.

2.1.2. Spall Growth Phase

The objective of this phase is to capture snapshots of the data measured by the sensors while growing the spall at a controlled pace. The protocol of this phase contains three stages that repeat each other until the end of the experiment. Figure 2 illustrates one cycle of this protocol, which includes three stages: the growth stage in blue, the monitoring stage in orange, and the collection stage in green. The black and purple vertical lines at the bottom of the graph indicate the load in each stage (black for 16 kN and purple for 6 kN). The vertical axis represents the normalized duration, which is the time duration normalized by the combined time of the monitoring and collection stages. The vertical line represents the shaft speed.

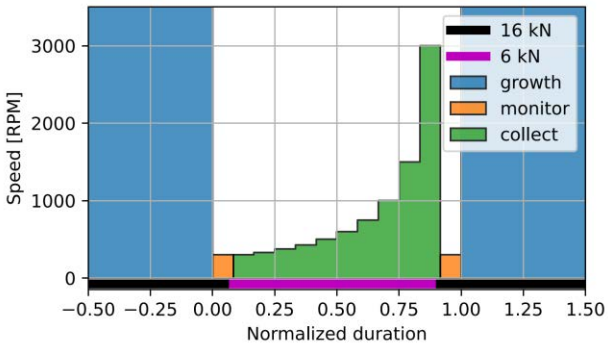


Figure 2: Example of one protocol cycle for spall growth. One cycle consists of three stages: 1: growth in blue, 2: monitor in orange, 3: collect in green.

In the growth stage, the aim is to accelerate spall growth; therefore, the bearing is subjected to a radial load of 16 kN, and the shaft rotational speed is set to 6000 RPM. One cycle of this stage lasts approximately 50 minutes.

In the monitor stage, two measurements are conducted at a high load with a speed of 300 RPM, one at the beginning of the collection stage and one at the end. In these measurements, the changes in the load cell are clearer and therefore will be used in this study for the load-based algorithm.

In the collection stage, measurements are taken from the sensors. In this phase, the load is reduced, and the speed changes to 10 different speeds spaced between 300 and 3000 RPM. The measurements at this stage will be used for future research. This stage takes around 20 minutes.

The experiment was halted approximately every 5 million revolutions for visual inspections. The test is stopped when a critical spall size, exceeding two times the distance between rolling elements, is reached. Beyond this size, two rolling elements no longer bear any load, which can lead to accelerated spall growth and, consequently, a high risk of critical failure.

3. Analysis of Load Cell Signal

In a faulted bearing, with a spall not larger than the distance between two rolling elements in the outer race, the interaction of the rolling element and the spall can be roughly divided into two stages. In one stage, none of the rolling elements interacts with the spall. The other stage is when one of the rolling elements is interacting with the spall; both stages are illustrated in Figure **3**.

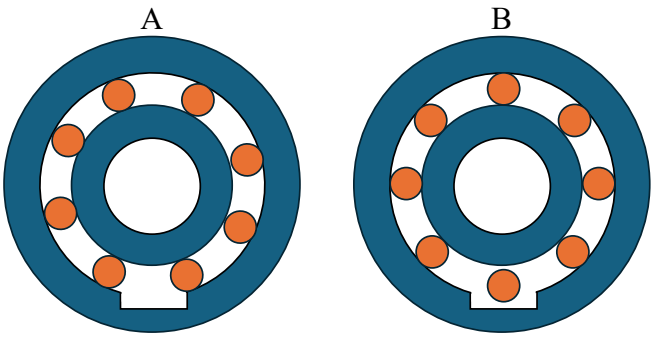


Figure 3: Illustration of REB interaction with outer race spall; (A) none of the RE interact with the spall (B) One RE enters the spall.

In stage one, the force applied to the bearing is divided among all the rolling elements in the bearing. In stage two, one of the rolling elements is inside the spall and, therefore, does not carry any load. This results in a different distribution of the load which appears to be observable on the load cell. By detecting these changes in the load cell, one can estimate the duration of the interaction with the spall, which can then be easily calculated to determine the spall length. At higher speeds, the transition between stages occurs more rapidly, meaning the system doesn't have enough time to stabilize, making it more challenging to detect in the time domain.

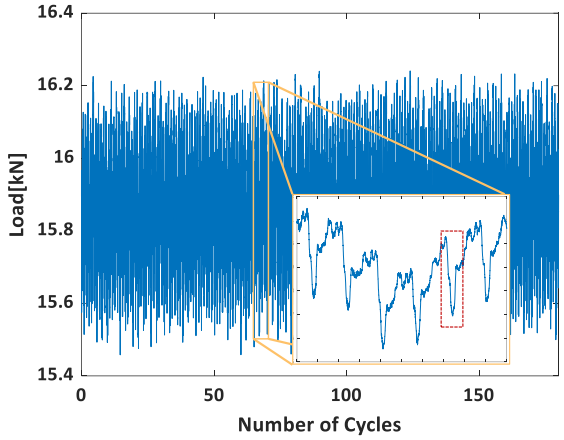


Figure 4: Raw load signal at 300 RPM with zoom on one cycle, equivalent to one shaft rotation: the area of interest marked by the red dashed rectangle indicate the area rolling element over the defect.

When examining the load signals acquired from the experiment at low speeds, it is possible to detect the interaction of the rolling element with the spall. Figure 4 shows an example of a signal with the visible interaction marked. To automate the process of identifying load distribution changes in the signal, a seven-step algorithm is proposed. The steps are described in Figure 6, with each step designed to emphasize and isolate the interaction of the rolling element with the spall. Each one of the steps is explained:

1. The load signal is detrended, by subtracting the "smoothed" signal from the original signal, making the signal centered around zero.
2. The interaction with the spall that occurs during the rotation of the shaft is periodic in the time domain when the speed is constant. However, even when setting the test rig to a constant speed, the speed is never truly constant. Therefore, angular resampling of the detrended load signal is conducted, converting the signal to the cycle domain.
3. Bearings are asynchronous components due to slippage (Sol et al., 2022). When employing Modified SA, as further explained in point 5, one can obtain a signal with isolated synchronous elements to the shaft's frequencies. By subtracting the SA signals from the original signal, the discrete shaft synchronous frequencies are removed, mitigating the interferences of other rotating components. This yields a signal containing only the asynchronous components. This algorithm is known as de-phase (Klein, 2017).
4. The cycle of interest is the interaction between the rolling element and the spall. Therefore, angular resampling is performed again based on the BPFO. Unlike the angular resampling based on the shaft's speed in step 2, the angular resampling in this step ensures a consistent number of samples in each cycle of the BPFO.
5. Modified Spectrum Analysis (MSA) (Koren, 2017) is utilized in this scenario. In MSA, the signal is segmented into N parts. The amplitudes of the Fourier Transforms (FT) for these segments are then averaged, resulting in an MSA signal with the averaged amplitude and phase information from a single segment. The fundamental steps of the MSA algorithm are outlined in Equations 1 and 2. Where N is the number of segments into which the signal is divided, and x_n represents a single segment of the signal. This technique is employed to isolate signal components asynchronous to the BPFO, including noise.

$$|\bar{X}| = \frac{1}{N}\sum_{n=1}^{N}|fft(x_n)| \quad (1)$$

$$MSA = ifft\{|\bar{X}| \cdot exp(j \cdot \angle fft(x_1))\} \quad (2)$$

6. A dynamic threshold is established by computing a percentage of the difference between the signal's highest and lowest points. Initially, a lower threshold is implemented for smaller spalls, which are more susceptible to noise interference. Once the estimated spall length reaches a predetermined value, a higher threshold is activated. This adjustment is intended to enhance accuracy. Figure 5 shows a processed signal with a set threshold indicated by a yellow dashed line.

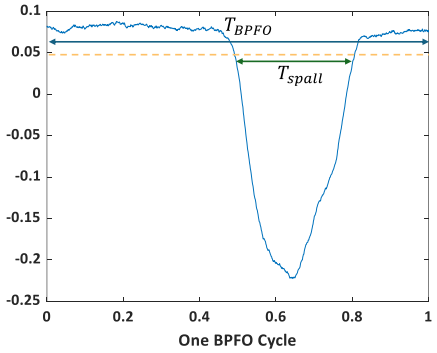


Figure 5: Processed signal with threshold indicated by a yellow dashed line; $\boldsymbol{T_{spall}}$ denotes the number of points below the threshold, and $\boldsymbol{T_{BPFO}}$ represents the total number of points in the MSA signal.

7. The determination of the pulse length involves calculating the percentage of values below the dynamic threshold. With the utilization of Equation 3, estimation of the spall size becomes feasible.

Here, RE represents the distance between two rolling elements, T_{spall} denotes the number of points in the synchronous average below the threshold, and T_{BPFO} represents the total number of points in the MSA signal.

$$S = RE \cdot \frac{T_{spall}}{T_{BPFO}} \quad (3)$$

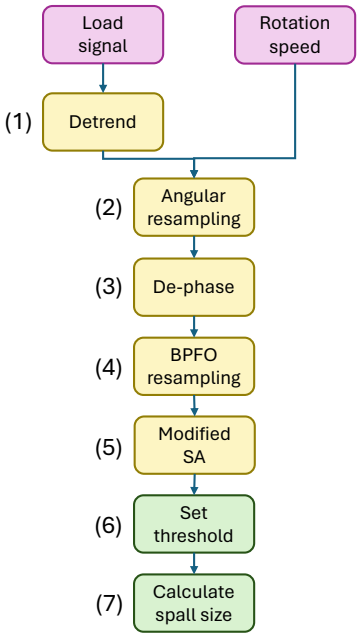


Figure 6: Block diagram of the load-based algorithm.

4. RESULTS

To validate the load algorithm, visual inspections were conducted during the endurance experiment. The tested bearing is a SKF N209 ECP cylindrical roller bearing. During each visual inspection, only the outer ring was disassembled, examined, and photographed. The spall size was measured by counting the number of pixels the spall occupies in each photo. An example from two visual inspections is shown in Figure 7. The spall lengths calculated from the proposed algorithm and the visual inspections were plotted for comparison and presented in Figure 8.

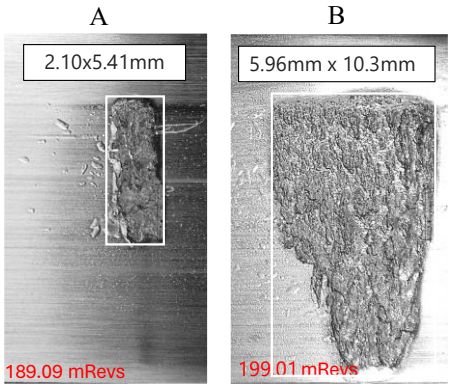


Figure 7: Visual inspections during endurance test: (A) at 189.09 million revolutions and (B) at 199.01 million revolutions. direction of the RE is from right to left.

It is evident that the estimated spall size by the load algorithm follows the trend of the measured sizes. However, in some cases, the estimated spall size deviates from that trend. These deviations sometimes occur after the visual inspections. The process of disassembling and reassembling could significantly alter the measured signals, as noted in a previous study (Heng et al., 2009). However, in the load signals, the impact is relatively small compared to acceleration and can be mitigated by using smoothing techniques. In other cases, the changes could be related to machinery malfunction, which contaminated the measurements. Despite these deviations, the suggested algorithm has shown good results and, in most cases, has been able to estimate the spall length accurately.

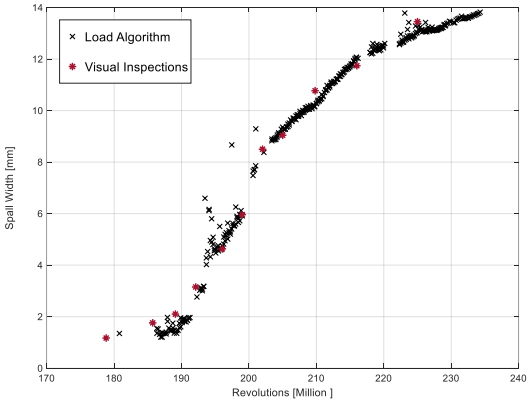


Figure 8: Comparison between the results of the load-based algorithm and the visual inspection.

In this work we present an algorithm to track spall size continuously in a robust manner in a lab environment by using existing load cells. Our proposed method can be used to validate spall size estimation algorithms. Moreover, it can be used to further study the physics of spall propagation, e.g., understanding the effects of speed and load.

5. CONCLUSION

In conclusion, bearings play a vital role in nearly all rotating machinery, highlighting the necessity of accurately estimating the severity of defects within them. As of today, there is no robust method for severity estimation in bearings, which can be used in all machinery. Endurance tests are crucial in bearing research, providing valuable insights into spall growth, and accurately labeling the data is essential for understanding this process.

Traditionally, labeling has relied on visual inspections during endurance tests, which can significantly alter vibration analysis results. This study introduces a load-based algorithm that eliminates the need for visual inspection, thus providing a more extensive dataset for labeling the severity of spalls. Although load cells are not typical components in machinery, they are common in experimental test rigs and can greatly assist with future research. The load-based algorithm was validated via visual inspection, demonstrating good agreement between the two methods. Not only does this algorithm streamline the testing process, but it also serves as a valuable tool for future studies, enabling researchers to track spall propagation and establish ground truth for developing acceleration-based algorithms. Overall, the implementation of this load-based algorithm represents a significant advancement in bearing defect analysis, offering improved labeling accuracy and opening up new avenues for further research in the field.

ACKNOWLEDGEMENT

Tal Bublil is supported by a scholarship sponsored by the Ministry of Science & Technology, Israel.

NOMENCLATURE

x_n	single segment of the signal
$\lvert\bar{X}\rvert$	average of the segment amplitudes
N	numbers of segments
MSA	MSA signal
RE	distance between two rolling elements
T_{spall}	number of points representing the spall
T_{BPFO}	total number of points in the MSA signal
S	spall length

REFERENCES

(ISO), I. S. O. (2016). *20816-1 Mechanical vibration - Measurement and evaluation of machine vibration - Part 1: General guidelines*.

Bechhoefer, E., & Schlanbusch, R. (2018). Calculating Remaining Useful Life in an Embedded System. *Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM*, 1–9.

Epps I K. (1991). *An investigation into vibrations excited by discrete faults in rolling element bearings.*

Gazizulin, D., Cohen, E., Bortman, J., & Klein, R. (2019). Critical Rotating Machinery Protection by Integration of a “fuse” Bearing. *International Journal of Critical Infrastructure Protection*, *27*, 100305.

Gebraeel, N., Lawley, M., Liu, R., & Parmeshwaran, V. (2004). Residual Life Predictions From Vibration-Based Degradation Signals: A neural network approach. *IEEE Transactions on Industrial Electronics*, *51*(3), 694–700.

Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating Machinery Prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724–739.

Klein, R. (2017). Comparison of Methods for Separating Vibration Sources in Rotating Machinery. *Mechanical Systems and Signal Processing*, *97*, 20–32.

N. Koren, I. Dadon,J. Bortman, R. Klein, (2017). Steps Towards Fault Prognostics of Gears. International Journal of Condition Monitoring, vol. 7, no. 1, pp.10-15.

Ma, L., Kang, J. S., & Zhao, C. Y. (2012). Research on Condition Monitoring of Bearing Health Using Vibration Data. *Applied Mechanics and Materials*, *226–228*, 340–344.

Madar, E., Galiki, O., Klein, R., Bortman, J., Nickell, J., & Kirsch, M. (2022). A New Model for Bearing Spall Size Estimation Based on Oil Debris. *Engineering Failure Analysis*, *134*(September 2021), 106011.

Malla, C., & Panigrahi, I. (2019). Review of Condition Monitoring of Rolling Element Bearing Using Vibration Analysis and Other Techniques. *Journal of Vibration Engineering and Technologies*, *7*(4), 407–414.

Medvedovsky, D., Ohana, R., Klein, R., Tur, M., & Bortman, J. (2022). Spall Length Estimation Based on Strain Model and Experimental FBG Data. *Mechanical Systems and Signal Processing*, *171*(February), 108923.

Moazen-ahmadi, A., & Howard, C. Q. (2016). A Defect Size Estimation Method Based on Operational Speed and Path of Rolling Elements in Defective Bearings. *Journal of Sound and Vibration*, *385*, 138–148.

Moazen Ahmadi, A., Howard, C. Q., & Petersen, D. (2016). The Path of Rolling Elements in Defective Bearings: Observations, Analysis and Methods to Estimate Spall Size. *Journal of Sound and Vibration*, *366*, 277–292.

Portal, O., Madar, E., Klein, R., Bortman, J., Nickell, J., & Kirsch, M. (2022). Towards Bearings Prognostics Based on Oil Debris. *Proceedings of the Annual*

Conference of the Prognostics and Health Management Society, PHM, *14*(1), 1–6.

Harris, T.A., & Kotzalas, M.N. (2006). Bearing Endurance Testing and Element Testing. (5th ed.) *Advanced Concepts of Bearing Technology: Rolling Bearing Analysis, Fifth Edition.* (763-792). Boca Raton.

Sawalhi, N., & Randall, R. B. (2011). Vibration Response of Spalled Rolling Element Bearings: Observations, Simulations and Signal Processing Techniques to Track the Spall Size. *Mechanical Systems and Signal Processing*, *25*(3), 846–870.

Smith, W. A., & Randall, R. B. (2015). Rolling Element Bearing Diagnostics Using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, *64–65*, 100–131.

Sol, A., Madar, E., Bortman, J., & Klein, R. (2022). Autonomous Bearing Tone Tracking Algorithm. *PHM Society European Conference*, *7*(1), 466–472.

Zhang, H., Borghesani, P., Randall, R. B., & Peng, Z. (2022). A Benchmark of Measurement Approaches to Track the Natural Evolution of Spall Severity in Rolling Element Bearings. *Mechanical Systems and Signal Processing*, *166*(March 2021), 108466.

Zhang, H., Borghesani, P., Smith, W. A., Randall, R. B., Shahriar, M. R., & Peng, Z. (2021). Tracking the Natural Evolution of Bearing Spall Size Using Cyclic Natural Frequency Perturbations in Vibration Signals. *Mechanical Systems and Signal Processing*, *151*, 107376.

BIOGRAPHIES

Tal Bublil is currently a Ph.D. student in the BGU-PHM LAB at the Department of Mechanical Engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. He completed his bachelor's degree with the highest honors in mechanical engineering at Ben-Gurion University of the Negev and completed his master's degree through the fast-track program.

Cees Taal has a research background with over ten years of experience in the field of sensor signal processing and machine learning. He has worked in academia (Delft University of Technology, Delft, The Netherlands, and KTH Royal Institute of Technology, Stockholm, Sweden) on audio and speech processing, after which he held various industrial positions in biomedical engineering (Philips Research, Eindhoven, Netherlands) and the energy domain (Eneco, Rotterdam, Netherlands). He is currently appointed as a Technologist of SKF, The Netherlands, where he is responsible for defining a research strategy in the field of bearing diagnostics and prognostics. Cees received the IEEE Signal Processing Society Best Paper Award in 2016.

Bert Maljaars is a researcher in SKF Research and Technology Development, working on diagnostics and prognostics in bearing condition monitoring. He received his MSc degree in Mechanical Engineering from Eindhoven University of Technology and afterwards his PhD degree (2017). His main research interests are state estimation, signal processing, physical modeling, optimization and control.

Renata Klein received her Ph.D. in the field of Signal Processing from the Technion, Israel Institute of Technology. For 17 years she managed the Vibration Analysis department in ADA-Rafael, the Israeli Armament Development Authority. Later, as the Chief Scientist in RSL- Electronics, she invented and led the development of a vibration based diagnostics and prognostics system that is used successfully in combat helicopters and UAVs of the Israeli Air Force. Renata is the CEO and owner of R.K. Diagnostics. In this role, and per invitation from Safran Aircraft Engines, she developed a full set of vibration based diagnostics and prognostics algorithms for jet engines. These algorithms are being integrated into the next generation of CFM jet engines. In recent years, Renata has focused on supervising academic research programs in the area of rotating machinery prognostics. Jointly with Prof. Jacob Bortman, she co-manages the BGU PHM Lab in Ben Gurion University of the Negev, teaches and provides supervision to MSc and PhD students.

Jacob Bortman is currently a full Professor in the department of mechanical engineering and the head of the PHM Lab in Ben-Gurion University of the Negev. Retired from the Israeli air force as brigadier general after 30 years of service with the last position of the head of material directorate. Chairman and member of several boards: director of business development of Odysight Ltd, Chairman of the board of directors, Selfly Ltd., board member of Augmentum Ltd., board member of Harel finance holdings Ltd., Chairman of the board of directors, Ilumigyn Ltd. Editorial board member of: “Journal of Mechanical Science and Technology Advances (Springer, Quarterly issue)”. Head of the Israeli organization for PHM, IACMM - Israel Association for Comp. Methods in Mechanics, ISIG - Israel Structural Integrity Group, ESIS - European Structural Integrity Society. Received the Israel National Defense prize for leading with IAI strategic development program, Outstanding lecturer in BGU, The Israeli prime minister national prize for excellency and quality in the public service - First place in Israel. Over 80 refereed articles in scientific journals and in international conference.

Landing Gear Health Assessment: Synergising Flight Data Analysis with Theoretical Prognostics in a Hybrid Assessment Approach

Haroun El Mir, Stephen King, Martin Skote, Mushfiqul Alam, and Simon Place

School of Aerospace, Transport and Manufacturing, Cranfield University, MK43 0AL Bedford, U.K.

H.el-mir@cranfield.ac.uk
S.P.King@cranfield.ac.uk
M.Skote@cranfield.ac.uk
Mushfiqul.alam@cranfield.ac.uk
C.s.place@cranfield.ac.uk

ABSTRACT

This study addresses a critical shortfall in aircraft landing gear (LG) maintenance: the challenge of detecting degradation that necessitates intervention between scheduled maintenance intervals, particularly in the absence of hard landings. To address this issue, we introduce a Performance Degradation Metric (PDM) utilising Flight Data Recorder (FDR) output during the touchdown and initial roll phases of landing. This metric correlates time-series accelerometer data from a Saab 340B aircraft's onboard sensors with non-linear response dynamic models that predict expected LG travel and reaction profiles across a set of ground contact cycles within a single landing. This facilitates the early detection of deviations from standard LG response behaviour, pinpointing potential performance abnormalities. The initiator of this approach is the Landing Sequence Typology, which systematically decomposes each aircraft landing into successive dynamic periods defined by their representative boundary conditions. What follows is the setting of initial parameters for the ordinary differential equations (ODE)s of motion that determine the orientation and impact responses of the most critical components of the LG assembly. Solving these ODEs with the integration of a non-linear representation of an oleo-pneumatic shock absorber model compliant with CS25 aircraft standards produces anticipated profiles of LG travel based on factors such as aircraft weight and speed at touchdown, which are subsequently cross-referenced with real accelerometer data, enhanced by video footage analysis. This footage is crucial for verifying the sequence of LG touchdowns and corresponding accelerometer outputs, thereby bolstering the precision of our analysis. Upon the conclusion of this study, by facilitating the early identification of LG performance deviations in specific landing scenarios, this diagnostic tool shall enable timely maintenance interventions. This proactive approach not only mitigates the risk of damage escalation to other components but also transitions main LG maintenance practices from reactive to proactive.

1. INTRODUCTION

Landing gear (LG) operational health is of paramount importance in ensuring aviation safety and optimising maintenance practices. Accurate assessment of LG component health can prevent catastrophic failures and reduce unscheduled downtime. Given the unique challenges posed by LG structural health monitoring (SHM)—arising from the use of high-strength, low-toughness materials in primary LG components, with relatively smaller critical crack propagation thresholds compared to the airframe—there is a compelling need for tailored monitoring approaches. A crucial constituent of LG SHM involves the monitoring of load, usage, and/or signs of crack initiation to estimate the remaining fatigue life of its monitored component/s. As a consequence, a prominent number of proposed LG health monitoring techniques rely on direct sensor placements, which can be intrusive, add weight, and increase the risk of error and maintenance requirements due to the introduction of said sensors. This study thereby addresses a prominent issue in the current LG integrity assessment approach followed by operators and MROs: the inability to detect LG degradation that requires intervention between scheduled maintenance intervals without the presence of hard landings. By inspecting touchdown and follow-up roll data at each landing cycle of the aircraft being monitored, we aim to remove the need for additional sensors. A Performance Degradation Metric (PDM) is being formulated, wherein the correlation of accelerometer time-series outputs with outputs from dynamic Ordinary Differential Equations (ODE)s of motion solved by Simulink models provides an indication of whether the LG's reaction profile was typical or deviant. This approach shifts the focus

from identifying issues like structural cracks and bearing wear to detecting abnormalities through deviations in dynamic performance from the models derived from a distinct set of conditions under which the aircraft interacts with the ground, incorporating shock absorber behaviour, aircraft mass, and impact speed. Awaiting identical conditions for comparison would necessitate an impractical volume of test data and landings. Therefore, this strategy focuses on assessing how and to what extent each of these variables impacts each of the main LG's performance during each landing.

Data for this study were collected using the Cranfield University Saab 340B aircraft, operated by the National Flying Laboratory Centre. This twin-engine turboprop, known as the National Flying Laboratory, has been customised to include specific experimental and teaching equipment to enhance its utility as a flying laboratory. The key modification vital for this study is the installation of an Ekinox-D: An INS sensor that offers orientation, heave, and centimeter-level position accuracy.

The rest of the paper is organized as follows: Section 2 delves into the traditional and contemporary methods of LG maintenance, discussing the shift from time-based strategies to real-time health monitoring, illustrated through various studies and the integration of progressive monitoring systems like fiber-optic sensors. In Section 3, we outline our methodology, emphasizing the integration of video footage, on-board sensor data, and dynamic modelling to analyse aircraft landing dynamics. Data collection techniques and the specific analytics used to extract and process this data are also detailed. Section 4 projects the future direction of our research, outlining the subsequent phases including sensor data analysis, structural dynamic response assessment, and the continuous development of our Performance Degradation Metric (PDM).

2. BACKGROUND

2.1. Traditional LG Maintenance Approaches

Traditionally, LG maintenance has leaned on time-based preventive strategies and Non-Destructive Testing (NDT) methods, including magnetic particle inspection, ultrasonic testing, and eddy current testing, as Schmidt (2008) notes. These conventional methods, applied during fixed maintenance intervals, often necessitate the disassembly of LG components for thorough inspection. In this context, the introduction of progressive monitoring marks a significant shift in maintenance paradigms. For instance, Kaplan et al. (1997) demonstrated the application of damage tolerance methods to extend the life of LG assembly subcomponents of a CASA 212 aircraft beyond their initial Safe-life design limits. By conducting loads, stress, and crack-growth analyses, they determined tailored inspection intervals. This approach underscores the potential of integrating damage tolerance principles to refine LG maintenance practices, paving the way for the adoption of landing profile-specific and load-adaptive health monitoring. Despite their intuitive approach and its success in extending the gear's service life, their methodology does not support real-time nor near-real-time assessment of LG health—a capability our current study seeks to develop. Importantly, while their approach contributes to extending the safe operational life of LG components, our project does not address direct estimations of life extension beyond set service limits, focusing instead on identifying and addressing immediate health concerns in operational conditions.

2.2. Advancements in Real-Time LG Health Monitoring

Building on these developments, recent advancements have shifted focus towards real-time LG health monitoring systems. These often involve the placement of sensors on critical LG components to monitor their condition during operation, such as that proposed by Zhang et al. (2018), who studied the placement of fiber-optic sensors on the outer tube weld of a LG assembly to capture weld crack signals. Further illustrating this trend, the EU-funded E-LISA project aims to develop an intelligent test facility for electro-mechanical LG, which will include PHM functionalities for the electrical brake system (De Martin et al., 2022). This project focuses on integrating sensors and monitoring systems into a novel LG design to enable condition-based maintenance. Similarly, Delebarre et al. (2017) contribute to the expanding landscape of sensor-based health monitoring with their development of a wireless monitoring system for lightweight aircraft LG, which uses pressure sensors and accelerometers to measure the mass distribution on each LG and monitor the shock during the landing phase. The system aims to provide real-time information to the pilot and maintenance personnel to improve safety and ease maintenance operations.

2.3. Data Analytics and Physics-Based Modelling in LG Health Monitoring

Integrating health monitoring systems into the LG architecture presents numerous challenges, such as coping with the harsh operational environment, managing the constraints on sensor placement, and ensuring the reliability of data transmission and analysis. These hurdles notwithstanding, the advancements in sensor technology and data analysis techniques offer promising pathways to surmount these obstacles, thereby enhancing the efficacy of aircraft LG health monitoring. In this vein, the work by (Bakunowicz & Rzucidło, 2020) presents an approach to detecting aircraft touchdowns using virtual sensing techniques by employing data from accelerometers mounted on structural parts of the airframe, utilising continuous wavelet transformation (CWT) to identify unique frequency signatures characteristic of LG touchdown. The CWT method, focusing on the detection of aircraft touchdowns with a high degree of precision, aligns closely with the

present paper's emphasis on optimising aircraft sensor output for LG health assessment. By extracting critical frequencies from accelerometers on-board during touchdown, our approach seeks to isolate and analyse pre-impact signatures, enhancing the precision of our health assessment metrics. Another pertinent reference in the context of virtual sensing is the work of Hsu et al. (2022) and its continuation by Chang et al. (2023), where they harness Flight Data Recorder (FDR) accelerometer outputs from a fleet of aircraft to detect early signs of exacerbated LG shimmy, thus indicating potential degradation that could require maintenance beyond scheduled intervals. Their study covers the taxiing phase before take-off and following landing, employing machine learning (ML) to link accelerometer readings with maintenance records across various LG components. They subsequently predict potential faults with almost 100% accuracy on almost all LG subcomponents used in training their ML model based on expert input and extensive data from landing cycle-based maintenance actions recorded on those specific LG components. Our study, while also utilising accelerometer data, extends the analysis to include longitudinal accelerations and converges specifically on the dynamics of landing impact and the subsequent short roll period, used in this case to include jumps and consequentially the Landing Sequence Typology approach which thereby defines non-linear response models representing their corresponding periods, for a CS25 aircraft.

The development of physics-based models for LG dynamics and health prediction has garnered significant attention in the field of LG SHM. These models aim to capture the interactions between various LG components and the forces they experience during operation (Schmidt, 2021). Recent studies have furthered this endeavour, focusing on high-fidelity dynamic modelling, synthetic dataset generation, and the advancement of prognostic algorithms for enhanced predictive accuracy. Wu, Gu, and Liu (2007) have notably developed a Nonlinear Model Predictive Control (NMPC) algorithm for semi-active LGs, utilizing Genetic Algorithms (GA). This method demonstrates an enhancement in LG performance by optimising the damping characteristics at touchdown, validated through drop tests that confirm the simulation model's accuracy. The GA-based NMPC approach effectively addresses the complex nonlinear dynamics of semi-active LGs, ensuring optimal performance despite constraints like the control valve's rate and magnitude limitations. In our approach, unlike the empirical validation possible through drop tests as utilised by Wu et al. (2007), we navigate the absence of a drop-test rig by emphasizing the integration of real-world operational data and physics-based models to refine our simulation accuracy further. This is in line with Krüger and Morandini's (2011) emphasis on the critical role of numerical simulation in LG dynamics assessment. Their research highlights the significance of modelling LG's dynamic response to various load excitations, underscoring the importance of a comprehensive understanding of LG dynamics for safety and performance. Finally, De Martin et al. (2022) present the development of the E-LISA iron bird, an innovative test facility for LG systems that includes PHM functionalities for the electrical brake system. The E-LISA project aims to reproduce the dynamic loads on the LG during landing, taxiing, and take-off, as well as the real contact between the LG wheel and runway. This approach aligns with our research objective of integrating real-world operational data and physics-based models to refine simulation accuracy and develop a hybrid approach for LG health assessment. De Martin. et al. (2022) present a high-fidelity dynamic model of the test rig, which incorporates the effects of runway-irregularities. This model serves as a foundation for generating synthetic datasets representative of various operating conditions and degradation levels, facilitating the development of prognostic algorithms. Their approach is similar to our use of physics-based models to predict the degradation of LG performance over time, and it highlights the importance of incorporating realistic operational conditions and representative component interactions in the set dynamic equations used to represent the conditions of a landing.

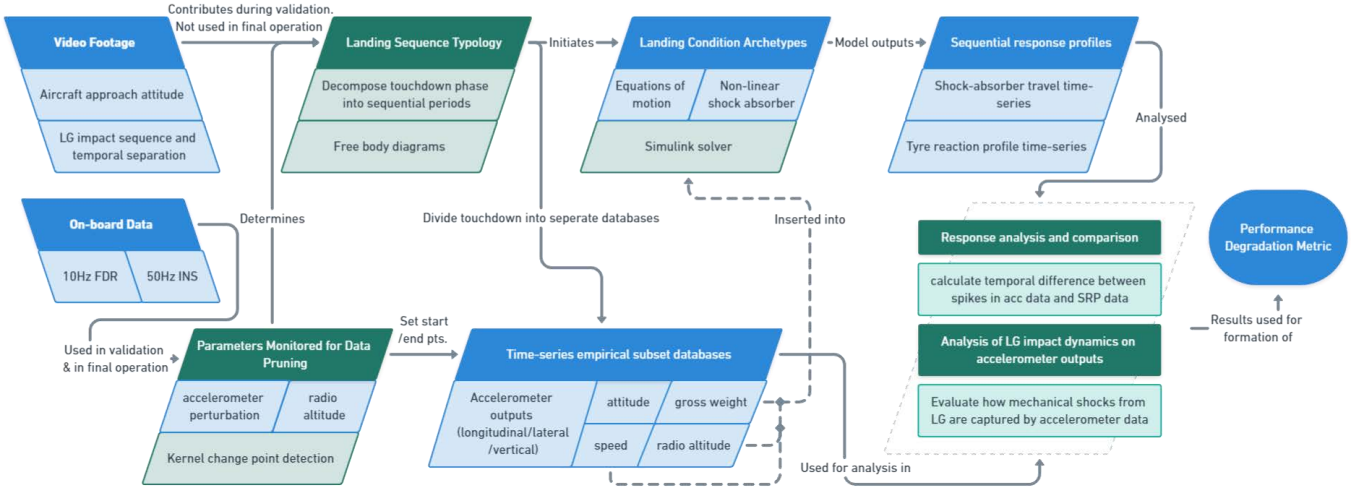


Figure 1. Integrated Framework for Aircraft Touchdown Analysis

3. METHODOLOGY

The methodology of this research is designed to analyse aircraft landing dynamics by integrating aircraft touchdown video footage, on-board sensor data, and bookcase non-linear response dynamic models, or 'archetypes', representative of the touchdown phases of each landing analysed. This multi-faceted approach allows for a robust examination of the impact sequences and a connection to the oleo-pneumatic shock-absorber (OSA) behaviours of the LG associated with different landing types. The study focuses on the following key aspects: capturing precise landing dynamics through video and sensor data, categorising landing types, formulating and solving ODEs to simulate these events, and validating these simulations against real-world data as feasibly as possible. Details follow in the subsections below, with corresponding visualisations provided in Figure (1), where the actions and outputs are denoted in green and blue blocks, respectively.

3.1. Data Collection

3.1.1. Video Footage Acquisition

A mirrorless APS-C video camera equipped with a telephoto lens is positioned on a fluid-head-equipped tripod by the runway border to record the final approach and touchdown. Operating at a frame rate of 29.97 fps and keeping the aircraft in-frame while extending the focal length to include only the undercarriage in the frame as soon as the aircraft is critically close to the airstrip, we ensure that each phase of the LG's contact sequence with the runway is meticulously documented. To ensure clarity and precision in the footage, the camera's shutter speed is set to at least four times the frame rate. This serves two critical purposes: it counteracts the shutter roll effect noticeable during fast panning—important for preventing deformations in the objects in-video, affecting important parameters such as adding distortions to tire deformation, which would be misleading—and it minimises motion blur to capture crisp imagery (when inspecting each frame in the video) of exact moments of touchdown, spin-up, spring-back, and hop. Additionally, the ISO setting is carefully controlled to prevent excessive photo grain, which impairs the accurate identification of the wheel edges contacting the airstrip. This footage is crucial for visualizing the aircraft's attitude at approach and touchdown, and the temporal separation between all undercarriage units; the main right, main left, and nose gear contacting the runway. The video data serves two primary purposes: it provides a visual reference for validating sensor data (temporal OSA impact delivery to on-aircraft accelerometer response output) and helps in identifying any discrepancies between observed and simulated main LG assembly behaviours. In Figure (2), an example of the footage contents may be seen.

Figure 2. touchdown footage frame

3.1.2. On-board Data

The aircraft is equipped with an IMU as part of a custom fit Inertial Navigation System (INS); the Ekinox-D, operating at sampling rates of 50Hz. The onboard data acquisition takes place by the use of the Curtiss-Wright/ ACRA Control KAM-500 system, which collects analog data from the Saab 340B's on-board sensors, including the Rockwell Collins AHS-3000 Attitude Heading Reference System. This setup captures essential aircraft dynamics and engine metrics using the Commercial Standard Digital Bus (CSDB) protocol (Alam, Whidborne, and Westwood, 2024). The data from these sensors are filtered to focus specifically on the touchdown phase, where detailed information about acceleration spikes and other dynamic responses is crucial for later analysis and simulation. The parameters recorded by these instruments include data on:

- Inertial Measurement Unit (IMU) and navigation: roll, pitch, heading, heave, surge, and sway from a MEMS (Micro-Electro-Mechanical Systems) sensor.
- Aircraft dynamics and engine metrics: accelerations, aileron and elevator deflections, angle of attack, fuel flow rates, gas generator speeds, propeller speeds, and turbine pressures.
- Environmental conditions: Airspeeds (indicated, true), Mach numbers, air temperatures, and radio altitudes.

In this study of aircraft dynamics, particularly before the initiation of gas generators and propellers, it is essential to calculate the root mean square (RMS) of accelerometer readings under stationary conditions. RMS is a statistical measure used extensively in signal processing to quantify the magnitude of a varying quantity. It provides a concise metric of the vibrational and transient accelerations experienced by the aircraft when it is static, which serves as a baseline for understanding the alterations in mechanical vibrations once the aircraft's propulsion components are activated. This baseline is critical for isolating and analysing the effects of mechanical and aerodynamic forces on the aircraft's structural integrity and operational efficacy. By calculating

the RMS value of accelerometer data while the aircraft is stationary, we can establish a reference point against which deviations caused by the gas generators and propellers can be measured, thereby offering insights into the dynamic behaviour of the aircraft under different operational conditions. Below are the RMS values which show minimal deviations and reaffirm the trustworthiness of the accelerometers for our use case:

INS MEMS Sensor:

- Lateral Acceleration: 0.0199g
- Longitudinal Acceleration: 0.0049g
- Normal Acceleration: 1.026g (indicative of gravity's influence)

Aircraft's on-board accelerometers:

- Lateral Acceleration: 0.0046g
- Longitudinal Acceleration: 0.0015g
- Normal Acceleration: 1.026g

3.1.3. Parameters Monitored for Data Pruning

In this step, a subset of the original time-series data is created based on the critical time period for analysis. Here, the Gaussian kernel, synonymous with the Radial Basis Function (RBF), is pivotal in the field of kernel-based change point detection (KPD), offering a nuanced approach to analysing complex data patterns. Its efficacy proves useful as a part of our method when filtering the time-series accelerometer readings for point-of-touchdown. This algorithm was rigorously tested across numerous flights, to ensure consistent touchdown indications across all accelerometer axes. Seeking a universally applicable method across diverse flight profiles, the single-point RBF approach (dynamic programming) was used. This method, applied to the derivative of time-series accelerometer readings showed promising adaptability and accuracy. Providing start and end points close to a chosen cut-off of radio-altitude also reduces its computing requirements and is currently the chosen approach. In Figure (3), you may see a plot of accelerometer measurements, their derivatives, and a red dashed line running vertically along the plot, indicating the KPD output corresponding to point of touchdown for the landing aircraft.

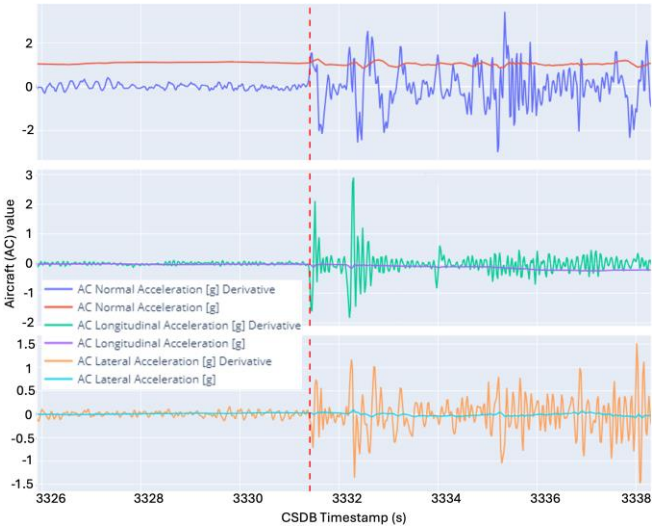


Figure 3. Accelerometer values and their derivatives w.r.t time for a level touchdown.

3.2. Landing Sequence Typology

To facilitate a structured analysis where causes and effects are recognised between landing load and landing variables, be they environmental, kinematics based, and/or temporal, each landing event is decomposed into several periods based on the amount of ground contact cycles. Each period is subsequently fitted to a category of distinct profiles based on observed dynamics and impact characteristics. The profiles are developed by analysing both video footage and sensor data to characterise each sequential landing period. This involves examining footage frames for the tyre impact timing, impact sequence, and the incidence angle, in addition to KPD-dictated touchdown indicators which serve in conjunction with the footage to dictate when the first period (linked to a profile) ends and the next begins. Each profile represents a set of initial conditions that are subsequently used to tailor the non-linear response archetypes. The profiles are categorised to be represented by, at their simplest:

- A smooth landing characterized by a negligible time difference between the touch-down of the rear right and left LG.
- High impact landings with minimal temporal separation between the rear LGs.
- Asymmetrical high impact landings affecting one side more than the other.
- Landings involving bounces, skips, or jumps.

By defining the characteristics of each period and linking it to a profile, the dynamic equations set for each profile can be adjusted to reflect the real-world dynamics observed during the data collection phase. This step ensures that they are representative of the variety of conditions the aircraft encounters in the duration of its single landing event.

3.2.1. Empirical Data Subsets Creation

Following the detailed decomposition of landing sequences as outlined in Section 3.2, and the rigorous data pruning mechanisms discussed in Section 3.1.3, the next phase focuses on compiling targeted time-series databases. These databases commence from the precisely determined touchdown point, leveraging the Gaussian kernel's efficacy in pinpointing this instant with high accuracy. The newly formed databases are confined to the parameters that are most indicative of landing dynamics and are crucial for the subsequent analysis:

- **Accelerometer Outputs**: Capturing the triaxial forces during the landing, these readings are pivotal for assessing the aircraft's response to touchdown dynamics.
- **Aircraft Attitude**: This includes the pitch, roll, and yaw of the aircraft at the point of touchdown, offering insights into the angular orientations that influence landing impacts.
- **Speed**: Ground speed and airspeed are included to correlate the velocity at touchdown with the landing impact severity.
- **Gross Weight**: The total weight of the aircraft influences the impact force and is thus critical for understanding the stress distribution on landing gears.
- **Radio Altitude**: For confirming the moment of touchdown and aids in synchronising other data streams.

Each database subset is tailored to represent a single impact cycle, which is identified based on the landing sequence typology. This approach ensures that each dataset is representative of specific landing conditions, thereby allowing for a more granular analysis of landing dynamics.

The speed, gross weight, and radio altitude are inserted into the completed landing condition archetypes for an output of the sequential response profiles that would allow for comparisons with the original subset databases containing the additional parameters representative of the period being inspected.

3.3. Landing Condition Archetypes

The preceding step, landing sequence typology, carries us closer to accurately representing the dynamics of a landing event by segmenting it into distinct sequential periods. Each period is tailored with specific boundary conditions corresponding to a respectively identified landing profile, enhancing the ground truth of our simulations, herein referred to as ‘archetypes’ which consist of non-linear dynamic ODEs combined with a model of a CS25 aircraft’s shock absorber and its interaction with the tyre and aircraft mass at level landing, which are critical for characterising the physical response of the aircraft's landing gear system under load. Given the lack of physical drop test rigs for empirical validation, it is imperative to assess the fidelity and robustness of these models.

Validation occurs in a bifurcated approach: Initially, the fidelity of the physics-based Simulink model is confirmed to ensure alignment between simulated performances with actual aircraft landing observations. This verification leverages detailed video stream analysis and FDR accelerometer data, which guide the establishment of stringent constraints and operational requirements specific to the landing gear system components in the simulation. These requirements are grounded in recognized benchmark methods, such as implementing damping strategies to mitigate resonance phenomena like shimmy and gear walk in the simulated landing gear assembly. A critical damping target, as stipulated by SAE International (2017) is reducing system oscillation to no more than a third of its original amplitude within three oscillation cycles post any perturbation.

3.3.1. Sequential Period Differential Equations

Using the data derived from the landing profiles, a set of ODEs is devised for each scenario. Free body diagrams (FBD) are utilised prior to forming these equations, ensuring that all relevant forces and interactions are accurately represented. The FBD of a level landing can be seen in Figure (4). These equations consider the mass, damping characteristics, and stiffness of the aircraft’s LG and structure. They include the non-linear characteristics of a CS-25 aircraft shock absorber, the interaction between the LG and the runway surface, and the effects of tyre dynamics on the LG system performance.

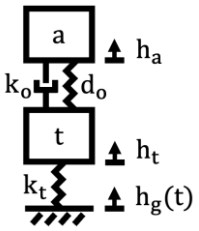


Figure 4. FBD for a level landing

The Simulink model in Figure (5) is adapted from that provided by (Jan R. Wright & E. Cooper, 2014). Simulink's environment allows for the continuous adjustment and real-time simulation of the equations, facilitating an iterative process of model refinement. The system is broken down into the aircraft rigid body mass and tyre mass, each with their own set of ODEs. The aircraft mass ODE includes terms for the spring and damper forces connecting the aircraft and tyre. The tyre mass ODE considers the forces from the OSA spring and damper, the tyre spring force, and runway height profile. A simple rigid aircraft landing system, assuming lift equals weight at touchdown, and ignoring spin-up and spring-back

and resulting LG motion due to them, is broken down as follows. Given:

- $\boldsymbol{h_a}$: Height of the aircraft mass from a reference point.
- $\boldsymbol{h_t}$: Height of the tyre mass from the same reference point.
- $\boldsymbol{h_g(t)}$: Runway height from the reference point, which is a function of time.
- $\boldsymbol{k_a}$: Spring constant connecting aircraft and tyre.
- $\boldsymbol{d_a}$: Damper constant connecting aircraft and tyre.
- $\boldsymbol{k_t}$: Spring constant connecting tyre and ground.

The resulting ODEs for the aircraft and tyre mass, respectively, are in Eq. (1) and Eq. (2) below:

$$m\ddot{h}_a = -k_a(h_a - h_t) - d_a(\dot{h}_a - \dot{h}_t) \qquad (1)$$

$$m_t\ddot{h}_t = k_a(h_a - h_t) + d_a(\dot{h}_a - \dot{h}_t) - k_t(h_t - h_g(t)) \qquad (2)$$

Additional ODEs are introduced for pitch and yaw dynamics depending on the period profile being modelled, considering the aircraft's moments of inertia, aerodynamic moments, and LG forces, and are a work-in-progress.

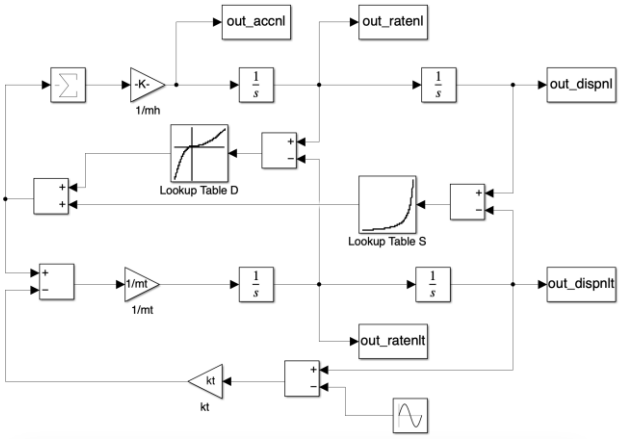


Figure 5. Simulink representation for a rigid-body level aircraft landing on main landing gear.

3.3.2. Non-linear Oleo-Pneumatic Shock Absorber

The OSA modelled employs a gas spring mechanism (the integral part affecting its dynamics), where the dynamics are significantly influenced by changes in gas volume and pressure during landing impacts. Its functionality is governed by the Ideal Gas Law, expressed as $PV^{\gamma} = C$, where P represents the absolute pressure, V the volume of the gas, γ the polytropic constant, and C a constant. The value of γ varies based on the OSA's operational conditions:

- Static Conditions (γ=1): This scenario represents steady, slow compressions such as during taxiing, where the temperature is maintained constant due to sufficient time for heat transfer.
- Dynamic Conditions (γ=1.3−1.4): During rapid compressions, such as landings, the process is adiabatic with no heat transfer, reflecting a higher γ value.

During the OSA's operation, as the LG encounters forces from the runway, the piston compresses, altering the gas volume. For a given change in volume ΔV caused by the piston stroke z, the new volume V_2 is given by $V_2 = V_1 - Az$, where A is the piston area. The corresponding pressures before and after compression, from the fully extended state V_∞ to the compressed state V_c are linked by Eq. (3):

$$P_\infty {V_\infty}^{\gamma} = P_c(V_\infty - Az)^{\gamma} \qquad (3)$$

The absolute pressure/displacement relationship can then be expressed in Eq. (4), where z_∞ is the fully bottomed distance (Jan R. Wright & E. Cooper, 2014):

$$\left(\frac{P}{P_\infty}\right) = \left(1 - \frac{z}{z_\infty}\right)^{-\gamma} \qquad (4)$$

According to Currey (1988), the typical characteristics for these calculations are as follows:

- Piston Area (A): Depends on the static pressure in the shock absorber, e.g., A=0.005m² if $P_{\text{static}} = 100$ bar.
- Pressures: $P_c = 3P_{\text{static}}$ and $P_\infty = 0.25P_{\text{static}}$.
- Volume Ratios: Assuming $V_\infty/V_c = 12$, then $V_\infty = V_c + A \cdot z_{\text{static}}$.

During landing, assuming that the lift equals the weight of the aircraft and neglecting tyre deformation to simplify the energy considerations, the kinetic energy of the aircraft equates to the work done by the OSA as in Eq. (5) (Jan R. Wright & E. Cooper, 2014):

$$\frac{1}{2}mv_y^2 = \eta_{\text{SA}}F_{\text{LGmax}}z_s = \eta_s\eta_{\text{LG}}Wz_s \qquad (5)$$

Where:

- m : Mass of half the aircraft plus part of the landing gear above the OSA.
- η_{SA} : Efficiency of the OSA, typically around 0.8.
- η_{LG} : LG load factor, ranging from 2 to 2.5 for CS-25 aircraft, representing the ratio of (static + dynamic reaction load) to (static load).
- W : Weight of the aircraft, equal to mg.

The force generated by the OSA, which is crucial for mitigating the impact during landing, is a function of the pressure differential across the piston. This force contributes to the overall dynamics of the aircraft's LG by opposing the landing load and dissipating kinetic energy. This is then translated into the Simulink environment through a series of blocks representing the aircraft's landing dynamics. The forces calculated from the OSA's pressure and volume changes are fed into the model to simulate the periods within the real-time landing event. These blocks use look-up tables generated from the aforementioned theoretical calculations.

3.3.3. Sequential Response Profiles

Sequential response profiles are derived from the outputs of the Simulink model to assess the performance of the OSA and the travel behaviour of the main LG during each sequential period. These profiles are essential for evaluating what similarities can be inferred between the archetypes and the empirical subset time-series data. The response profiles include the shock-absorber travel time-series, which tracks the displacement and normalised load absorbed over time, and the tyre reaction time-series, documenting the reaction forces of the tyre which reflect the dynamics of the unsprung mass. The analytical approach involves aligning the data starting at the moment of touchdown, identified by radio altitude and verified through accelerometer data, ensuring that the simulation phases are synchronized with the actual event timings. The Simulink solver continuously processes the differential equations representing the landing dynamics. The shock-absorber's travel and tyre reaction forces are methodically captured and plotted to provide an examination of the forces at play during the landing.

3.4. Comparison with On-board Data

In parallel, while video footage is used to validate the temporal and sequential accuracy of the archetypes in some capacity, the sequential response profiles (Simulink outputs) are compared to the time-series empirical accelerometer output corresponding to each of these periods. In our study, the primary objective of comparing Simulink model outputs to empirical accelerometer data is to establish a robust relationship in terms of observed trends and to correlate these observations with specific landing profiles, such as a hard level landing. This analysis involves comparisons of both Simulink outputs and accelerometer data collected from the aircraft during defined landing scenarios. The goal is to systematically expand this analysis across multiple flights and varying initial conditions, thereby compiling a comprehensive set of correlations between the model's predictions and the actual accelerometer responses recorded on the aircraft. For each period of each landing event analysed, the model outputs and accelerometer readings are compared to determine how closely the simulated responses (from the Simulink model) align with the real-world data under similar operational conditions. Key parameters considered during these comparisons include aircraft speed, gross weight, and radio altitude variation which would give us vertical speed at the point of touchdown. Through repeated evaluations across diverse flight conditions, this method allows us to refine our understanding of the dynamic interactions between the aircraft's LG and the runway surface.

3.5. Performance Degradation Metric Definition

As the dataset grows, encompassing a wider array of flight profiles, we progressively build a Performance Degradation Metric (PDM). This metric is designed to assess, using only the time-series output from the aircraft's accelerometers at touchdown, whether the observed accelerometer responses align with expectations derived from our simulations and previous correlations. This involves two critical analyses: first, evaluating the output of the Simulink model corresponding to the given profile (in the form of sequential response profiles for the specific period), and second, examining the established relationships between key accelerometer performance indicators, including peak-to-peak time, temporal peak separation, and time interval analysis relative to specific thresholds, and their alignment with Simulink model outputs. Based on the discrepancies identified between the simulated results and the actual data, adjustments are made to the ODEs and their parameters in the Simulink model. These adjustments may include changes in the damping coefficients, stiffness parameters, and mass distribution within the landing gear system. Each iteration aims to reduce the error margin and enhance the fidelity of the model. This approach aims to ensure as much as possible that each phase of the investigation contributes to a systematic and scalable understanding of the landing dynamics, which is crucial for advancing the predictive capabilities of our models.

Central to the separation in terms of model comparison of this analysis is the delineation of the minimal interval necessary for both main LGs to contact the runway simultaneously in a level touchdown—a scenario that equally distributes the landing load but remains exceedingly rare due to the imperative for pilots to adjust for crosswinds through controlled bank angles and the inherent inconsistencies present in airstrip surfaces. In recognising that aircraft landings may encompass a complex combination of the aforementioned scenarios, the PDM shall incorporate a nuanced measurement of the intensity and category of each phase encountered, leading to the point of analysing probability of performance degradation; assessing each LG unit's potential for operational wear (be it the right or left LG assembly). By continuously refining the correlation between simulated outcomes and actual flight data, our study aims to provide reliable predictive tools that can effectively anticipate operational degradation of the aircraft's landing systems under varied operational conditions.

This PDM is to output a relative operational health status of the main LG assemblies as shown in Figure (4). This plot displays the relative operational health status of the main LG assemblies over the course of successive landings. The initial operational health status is set at 100% at the commencement of operation (0 landings), with the Safe-life indicating the theorised lifespan, shown as a fixed endpoint in the plot at a landing life of 60k. The plot simplistically portrays the relative operational health as declining linearly; however, this does not reflect real-world conditions and is merely a simplification for illustrative purposes. The plot serves as a theoretical model, illustrating the projected outcomes we aim to achieve by the conclusion of the project. Key components include:

- **Safe Life Health Status**: The dashed red line serves as a theoretical performance threshold. Should the operational health of any LG assembly drop below this line, as predicted by the hybrid model, this would suggest potential risks at which an inspection is required.
- **Left and Right LG Hybrid Approach Health Status**: The blue and green lines show actual health status tracking for left and right LG, respectively, with maintenance actions represented by 'x' markers.
- **LG Failure at 5400 Landings**: This trend exemplifies the characteristic decline preceding a failure event.
- **LG Health in Ideal Low Wear Conditions**: A trend representing a LG assembly that has undergone extremely low-impact landing cycles.

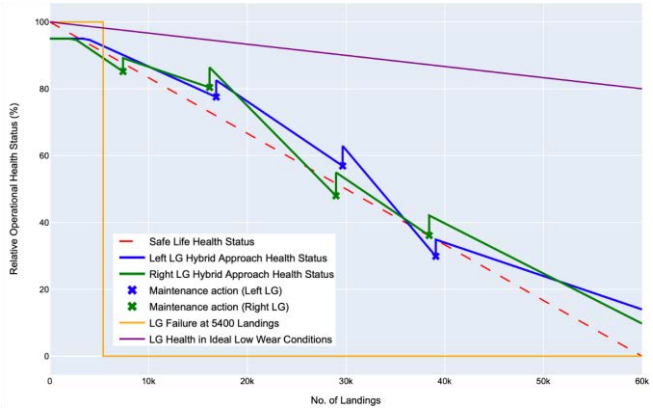


Figure 6. LG Operation Health Status

The value of the relative operational health status represents the current operational condition of the system, rather than direct LG part degradation. Its value is relative to the corresponding value of the Safe Life Health Status at that no. of landings. In Figure (7), a closer examination of the initial segment of the plot in Figure (6) reveals inherent uncertainties in the model's operation, stemming from the requisite number of landing cycles needed to establish reliability. Currently, this figure is illustrative and subject to refinement as our project evolves towards more precise and realistic estimations.

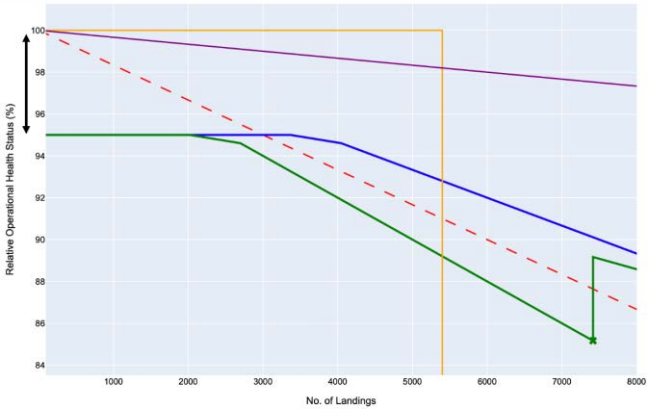


Figure 7. A close-up on no. of landings required for model validity

4. PROJECT DIRECTION AND FUTURE WORK

This paper marks the commencement of a structured approach for enhancing LG health assessments by means of virtual sensing combined with landing scenario-representative empirical models. While this paper discusses the initial stages of the first study, subsequent planned investigations will further this exploration:

Study 1 - Sensor Data Analysis and LG Dynamics: This segment focuses on extracting and analysing data from the FDR and IMU, comparing these to LG response profiles that are a result of landing condition archetypes to detect deviations in accelerometer oscillations and other critical parameters. Objectives include:

- **Operational Condition Analysis**: Examining variations in LG dynamics across different operational conditions.
- **Performance Pattern Identification**: Identifying desirable performance patterns and recognising limitations.

Study 2 - Sensor Placement and Data Precision: This study aims to compare IMU and on-aircraft accelerometer outputs during the landing's touchdown and roll phases, to identify the most effective sensor placements for LG response evaluation. This assists in pinpointing LG performance patterns during crucial phases. The focus areas are:

- **Sensor Output Comparison**: Crafting strategies for comparing sensor outputs to underline strategic placement.
- **Filtering Techniques**: Applying filtering methods to sensor outputs for improved data accuracy.

Study 3 - Structural Dynamic Response: Initiates a quantitative examination of modal frequencies and structural resonances before landing, employing high-fidelity spectral analysis to differentiate these from frequencies observed

post-touchdown. This study encompasses high-fidelity spectral analysis to separate pre-impact from post-impact frequencies.

Results and Future Directions: Following these studies, we shall present:

- **PDM Development**: A more detailed discussion on the development and validation of the PDM, including an assessment of operational degradation in the port and starboard LG relative to maintenance schedules.
- **Empirical and Theoretical Insights**: A comparative analysis offering essential insights from our empirical data and theoretical models.
- **Case Studies**: Application of our hybrid approach to real-world scenarios.

Future initiatives will broaden these methodologies to encompass more aircraft components and scenarios, aiming to reduce aircraft downtime and enhance safety across various models.

REFERENCES

Alam, M., Whidborne, J., & Westwood, M. (2024). Navigation Algorithm for a Twin-Engine Turboprop Aircraft using an Extended Kalman Filter. *EuroGNC 2024 Forum*, Bristol, UK [to be published]

Bakunowicz, J., & Rzucidło, P. (2020). Detection of Aircraft Touchdown Using Longitudinal Acceleration and Continuous Wavelet Transformation. *Sensors*, *20*(24), 7231. https://doi.org/10.3390/s20247231

Chang, Y.-J., Hsu, H.-K., Hsu, T.-H., Chen, T.-T., & Hwang, P.-W. (2023). The Optimization of a Model for Predicting the Remaining Useful Life and Fault Diagnosis of Landing Gear. *Aerospace*, *10*(11), 963. https://doi.org/10.3390/aerospace10110963

Currey, N. S. (1988). *Aircraft landing gear design: Principles and practices*. American Institute of Aeronautics and Astronautics.

De Martin, A., Jacazio, G., & Sorli, M. (2022). Simulation of Runway Irregularities in a Novel Test Rig for Fully Electrical Landing Gear Systems. *Aerospace*, *9*(2), Article 2. https://doi.org/10.3390/aerospace9020114

Delebarre, Grondel, Dupont, Rouvarel, & Yoshida. (2017). Wireless monitoring system for lightweight aircraft landing gear. *2017 International Conference on Research and Education in Mechatronics (REM)*, 1–6. https://doi.org/10.1109/REM.2017.8075230

Hsu, T.-H., Chang, Y.-J., Hsu, H.-K., Chen, T.-T., & Hwang, P.-W. (2022). Predicting the Remaining Useful Life of Landing Gear with Prognostics and Health Management (PHM). *Aerospace*, *9*(8), 462. https://doi.org/10.3390/aerospace9080462

Jan R. Wright, & E. Cooper, J. (2014). Ground Manoeuvres. In *Introduction to Aircraft Aeroelasticity and Loads* (pp. 337–365). https://doi.org/10.1002/9781118700440.ch15

Kaplan, M. P., Willis, T., & Wolff, T. A. (1997). *Damage Tolerance Assessment of CASA Landing Gear*. 972626. https://doi.org/10.4271/972626

Martin, A. D., Jacazio, G., Parisi, V., & Sorli, M. (2022). Prognosis of Wear Progression in Electrical Brakes for Aeronautical Applications. *PHM Society European Conference*, *7*(1), 329–337. https://doi.org/10.36001/phme.2022.v7i1.3353

SAE International. (2017). *AIR4894: Landing Gear Stability*. SAE International. https://saemobilus.sae.org/content/AIR4894/

Schmidt, R. K. (2008). Monitoring of aircraft landing gear structure. *The Aeronautical Journal*, *112*(1131), 275–278. https://doi.org/10.1017/S0001924000002220

Schmidt, R. K. (2021). *The Design of Aircraft Landing Gear*. SAE International.

Wu, D., Gu, H., & Liu, H. (2007). GA-Based Model Predictive Control of Semi-Active Landing Gear. *Chinese Journal of Aeronautics*, *20*(1), 47–54. https://doi.org/10.1016/S1000-9361(07)60006-5

Zhang, H., Wang, S., & Yang, Q. (2018). Research of Monitoring the Landing Gear Damage Based on the Optical Fiber Acoustic Emission Technology. *IOP Conference Series: Materials Science and Engineering*, *452*, 022091. https://doi.org/10.1088/1757-899X/452/2/022091

Large Language Model-based Chatbot for Improving Human-Centricity in Maintenance Planning and Operations

Linus Kohl[1,2], Sarah Eschenbacher[1], Philipp Besinger[1] and Fazel Ansari[1,2]

[1]*Fraunhofer Austria Research GmbH, Center for Sustainable Production and Logistics, Vienna, 1040, Austria*
linus.kohl@fraunhofer.at
sarah.eschenbacher@fraunhofer.at
philipp.besinger@fraunhofer.at

[2]*TU Wien, Research Group of Production and Maintenance Management, Vienna 1040, Austria*
fazel.ansari@tuwien.ac.at

ABSTRACT

The recent advances on utilizing Generative Artificial Intelligence (GenAI) and Knowledge Graphs (KG) enforce a significant paradigm shift in data-driven maintenance management. GenAI and semantic technologies enable comprehensive analysis and exploitation of textual data sets, such as tabular data in maintenance databases, maintenance and inspection reports, and especially machine documentation. Traditional approaches to maintenance planning and execution rely primarily on static, non-adaptive simulation models. These models have inherent limitations in accounting for dynamic environmental changes and effectively responding to unanticipated, ad hoc events.

This paper introduces a *maintenance chatbot* that enhances planning and operations, offering empathetic support to technicians and engineers, boosting efficiency, decision-making, and on-the-job satisfaction. It optimizes shift scheduling and task allocation by considering technicians' skills, physical stress, and psychological state, thus reducing cognitive stress. The approach ultimately improves human performance and reliability, embodying a human-centricity in the domain of maintenance and health management.

The practical impact of the *maintenance chatbot* is illustrated through its application in maintenance of railway cooling systems. The presented use case demonstrates the chatbot's potential as a transformative tool in maintenance management. Finally, the paper discusses the theoretical and practical considerations, in particular in the light of regulative frameworks such as EU AI ACT, highlighting the future pathways for complying with responsible AI requirements.

1. INTRODUCTION

The industrial landscape is currently facing a significant challenge due to the shortage of skilled labors, exacerbated by the increasing complexity of machinery and technological systems, as well as green transition, leading to limiting production by 28% in the European Union (EU) (European Commission 2023). This shortage poses a critical threat to the operational efficiency and sustainability of maintenance operations within various sectors. The complexity of modern machines requires a high level of expertise, yet industries often find themselves compelled to hire workers who may not fully meet these competency requirements (Shin et al. 2021). The European Union estimates the investment needed to reskill and upskill in manufacturing to 4.1 billion EUR up to 2030 (European Commission 2023). This gap between the required and available skill sets leads to inefficiencies, increasing human failure, thus reducing reliability and increasing downtime, and a greater potential for errors in maintenance operations.

Simultaneously, advancements in GenAI and semantic technologies have opened new avenues for capturing and leveraging the domain knowledge of experienced professionals (Abu-Rasheed et al. 2024), and at the same time assisting them on improving their problem-solving capabilities, e.g. through query-answers with chatbots (Kohl und Ansari 2023b). These technologies, particularly Large Language Models (LLMs), demonstrate an unparalleled capacity to analyze and interpret complex datasets, including technical documentation, maintenance logs, and operational reports (Birhane et al. 2023). Their ability to generate contextually relevant, accurate responses based on vast amounts of textual information marks a significant step forward in the development of cognitive assistants for maintenance tasks.

The intersection of skilled labor shortages (as a problem space) and GenAI technologies (as a solution space) underscores a critical need for tools that can bridge the gap between the complexity of modern machinery and the competencies of the available workforce. Cognitive assistance in maintenance, facilitated by AI-driven solutions, offers a promising approach to address this challenge (Kohl und Ansari 2023a). By providing real-time, tailored information and support, such tools can enhance decision-

making, reduce cognitive load, and improve the efficiency of maintenance technicians who may not possess the full spectrum of required competencies and experiences. Furthermore, the integration of GenAI and semantic technologies in maintenance operations enables the preservation and dissemination of expert knowledge, mitigating the risk of knowledge loss due to workforce turnover or the retirement of seasoned professionals (Alavi et al. 2024). This capability is particularly valuable in light of the increasing complexity and specificity of modern industrial systems, where the loss of domain-specific knowledge can have significant operational impacts (Ansari 2019).

The need for cognitive assistance in maintenance is not only a response to the skilled labor shortage but also a strategic investment in the quality and reliability of maintenance operations. By enhancing the capabilities of maintenance technicians, engineers and planners, AI-driven tools can contribute to more resilient, efficient, and effective maintenance practices. The development and implementation of such tools, as exemplified by the LLM-based maintenance chatbot presented in this paper, represent a forward-looking approach to addressing the challenges of the contemporary industrial maintenance landscape (Romero und Stahre 2021). The following paper addresses the challenge of improving the workflow of maintenance operations and planning by leveraging LLM and semantic information.

The rest of the paper is structured as follows: In Section 2, the state-of-the-art is described, focusing on cognitive assistance system, Generative AI, especially Large Language Models. Thus, the research gap is identified. Section 3 introduces the system architecture and modular chatbot design, and Section 4 elaborates on its use case. Finally, Section 5 discusses the key findings and identifies the pathways for future research.

2. STATE OF THE ART

This section explores the capabilities and applications of cognitive assistance systems within industrial manufacturing, emphasizing their role in augmenting human capabilities. It highlights how these systems utilize advanced technologies such as LLMs and KGs to optimize task execution. Additionally, it addresses the implications of the EU AI Act, which mandates transparency and safety in the deployment of such AI-driven systems, ensuring their responsible application in industrial environments.

2.1. Cognitive assistance system

Digital assistance systems (DAS) support workers in production, assembly and logistics to carry out their tasks efficiently in line with the situation and context (Ansari et al. 2020). These systems facilitate tasks ranging from scheduling and information retrieval to more complex operations, leveraging user inputs to deliver relevant outcomes and insights (Pokorni und Constantinescu 2021). Cognitive assistance systems (CAS), particularly within the manufacturing sector, extend this concept by focusing on augmenting human capabilities in intricate tasks rather than substituting human efforts (Kernan Freire et al. 2023), which can draw conclusions from its experience on the basis of significant portions of suitably presented knowledge so that it provides more appropriate, accurate or up-to-date information in its next use. These systems are engineered to support complex activities, including lifelong learning (Freire et al. 2023), machine operation, and task execution, through advanced methods of human-machine interaction (Listl et al. 2021). Employing a broad spectrum of techniques such as natural language processing (NLP) (Ansari et al. 2021), pose estimation for ergonomic risk identification (Kostolani et al. 2022), perception, and augmented reality (Zigart und Schlund 2020), CAS are designed to foster an intuitive and efficient interface for users.

CAS, utilizing NLP for natural language understanding, generation, and dialogue management, represent the most widespread interaction modality within CAS (Kang et al. 2020). These CAS are capable of engaging users in meaningful conversations, thereby facilitating labor-intensive tasks across multiple sectors, including customer service, healthcare, education, and manufacturing, through efficient and reliable communication (Eloundou et al. 2023).

In the industrial context, the application of CAS is an evolving research domain with significant potential benefits (Mark et al. 2021). These include providing centralized access to diverse information systems, decision making (Rožanec et al. 2022), delegating tasks (Burggräf et al. 2021), and enabling hands-free and gaze-free interactions (Romero und Stahre 2021), thereby enhancing operational efficiency and safety. Additionally, CAS in manufacturing can serve as valuable tools for on-the-job training (Wang et al. 2022) and real-time machine parameter adjustments, thereby contributing to the flexibility and adaptability of manufacturing processes (Zheng et al. 2022). Such applications highlight the transformative potential of cognitive assistants in augmenting human work, optimizing task execution, and facilitating continuous learning and adaptation in complex industrial environments.

2.2. Generative AI and Large Language Models

According to the OECD, GenAI “creates new content in response to prompts, offering transformative potential across multiple sectors such as education, entertainment, healthcare and scientific research”(OECD Artificial Intelligence Papers 2024). It, therefore, significantly broadens AI's application spectrum (Gozalo-Brizuela und Garrido-Merchan 2023). At the heart of GenAI's advancements are LLMs like Generative Pre-trained Transformers (GPT), which have dramatically enhanced AI's language processing and generating

capabilities, offering applications from automating documentation to improving decision-making in industries.

Retrieval-Augmented Generation (RAG) (Jing et al. 2024) extends LLMs by integrating them with information retrieval systems, enabling real-time access to extensive databases for more precise, context-specific causal outputs (Zhou et al. 2024). This is particularly valuable in manufacturing and maintenance, where accessing up-to-date technical and diagnostic information is crucial (Kernan Freire et al. 2023).

AI agents represent a further advancement, capable of autonomous decision-making based on environmental learning and adaptation (Zhao et al. 2023). In the context of manufacturing and maintenance, these agents can autonomously monitor system health (Han und Tao 2024), predict (Saboo und Shekhawat 2024) and automate maintenance tasks (Sun et al. 2024), thereby reducing downtime and maintenance costs. It can therefore be said that current approaches can achieve relevant results through their purely probabilistic transformer architecture by using attention with classical RAG, but cannot use factual, linked knowledge.

2.3. Knowledge Graph

Knowledge Graphs (KGs) structure knowledge in graphs, connecting entities and their relationships, thereby facilitating semantic searches and data integration (Fensel et al. 2020). In GenAI applications, KGs enhance RAG (Zhu et al. 2024) by providing structured, semantically linked domain information to improve response accuracy and contextual relevance (Agarwal et al. 2020), particularly valuable in domain-specific applications like manufacturing (Yu 2022). KG therefore enhance capabilities of chatbots by providing them with structured context information on specific user requests (Li et al. 2021). By leveraging the rich semantic relationships within KGs, chatbots are able to understand and process user queries more effectively, navigating through complex information networks to retrieve or infer accurate answers (Yu 2021).

Within manufacturing, KGs encapsulate domain knowledge and causal relationships between failure modes and solutions, informed by Failure Modes and Effects Analysis (FMEA) (Razouk et al. 2023). This structured knowledge aids RAG systems in querying precise information for predictive maintenance and decision support, thereby streamlining maintenance protocols and diagnosing machinery issues through an understanding of causal links.

The synergy between KGs and RAG significantly enhances manufacturing operations' efficiency by enabling access to detailed domain knowledge, reducing downtime, and guiding accurate maintenance decisions, thus enhancing operational reliability and performance (Ansari et al. 2023).

2.4. EU AI Act

In response to the rapid developments in the field of AI in recent years, the European Union has implemented a regulatory framework for development, market introduction and deployment of AI-driven products, services, and systems. The framework is designed to guarantee transparency, accountability, and safety for both current and forthcoming AI technologies within the EU. Especially in the area of manufacturing a responsible application of AI is essential to mitigate risks and deliver business benefits (Besinger et al. 2024).

Since current pre-trained LLMs like the GPT-models (Brown et al. 2020) or Metas Llama-Series (Touvron et al. 2023) are trained outside of the European Union, the EU AI Act addresses this issue by extending its scope to include providers operating within the EU as well as those in third countries, particularly when the output of their AI systems is utilized within the Union. The EU AI Act defines different categories from no risk to high-risk. The use of AI in human interaction, emotion recognition, and content generation is categorizes as low risk (second category). Article 52 (European Commission 2024) addresses the regulatory requirements for providers and users (excluding end-users) of AI systems categorized as low risk. There are three critical areas pertinent to the case presented in this paper: Transparency in AI interactions, the Marking of synthetic content, and the Disclosure requirements for emotion recognition and biometric categorization.

Firstly, concerning Transparency in AI interactions, the legislation mandates that AI systems engaging in human interaction must inform users of their non-human nature, except in contexts where such interaction is inherently apparent. Secondly, the requirement for marking synthetic content, such as audio, images, videos, or text, created or significantly altered by AI there must be machine-readable marks signifying its artificially generated or manipulated status, except for minor edits. Lastly, concerning emotion recognition and biometric categorization, users must be informed about these processes, with data handling needing to comply with EU regulations (European Commission 2024).

2.5. Research Gap

In industrial maintenance, the accessibility and quality of critical data is a crucial issue. Despite the increasing availability of information from maintenance reports, personnel documents, and enterprise resource planning (ERP) systems, the effective use of this data remains largely untapped. Therefore, to the best of the authors' knowledge, current research has not sufficiently explored the use of LLM in chatbots in industrial applications, especially the use of linked data and documents in an agent network. This paper presents a novel way to combine LLM with RAG and KG in an intent-driven agent framework, providing a flexible,

generalizable and scalable approach for industrial maintenance.

3. METHODOLOGY

In the following the design of the system architecture, which enables an LLM-based maintenance chatbot is described. Further, we propose a modular agent layout for the chatbot. The architecture is based on by RAMI 4.0 (DIN 91345) and inspired by (Margaria und Schieweck 2019).

3.1. System Architecture

The system architecture for the application of an LLM-based maintenance chatbot, see Figure 1. is structured into three distinct tiers, namely data tier, analytics tier and presentation tier, each with specific components, capabilities and information flows designed to interact seamlessly within the broader ecosystem of the industrial application.

Data Tier: The *Event Broker* facilitates communication between the analytics components and the data sources. It manages the flow of real-time data to the Data Analytics (Stream) and routes information to and from the Prescriptive Analytics. The *Database* stores historical data, such as CAD-models, maintenance reports or technical data, which is subsequently used for trend analysis and informing predictive models. It also serves as a repository for collected data over time and connects them through semantic similarities, which leverages the suitability of natural language interaction. Vectorized data schemas in the Data Tier allow for efficient data retrieval. *Edge Devices* are directly connected to the database and serve as intermediaries between the physical sensors and the system's core data infrastructure. They perform preliminary data processing, filtering, and aggregation tasks.

Sensors, either attached to machines or environmental sensors, collect data about the operational status, health, and performance of the machinery and environmental status. This data is crucial for monitoring and maintenance purposes. It incorporates different database structures. For processing natural language, the core components are a vector database and a KG, which serve as the foundation for an efficient RAG pipeline. While vector databases enable efficient data retrieval through vectorized representation of domain specific data (Jing et al. 2024), KGs provide structured representation of the data (Pan et al. 2024). A combination of these components is leveraged to reduce hallucination and utilize information which is not inherent to the LLM. The KG, see Figure 1, enables the connection of ERP data with task and competence relevant information. This data model allows a holistic view on the maintenance process as well as the possibility for downstream agents for interconnected reasoning. *Machines* are the physical hardware being monitored and maintained. Connected to sensors, they are the source of the operational data fed into the system for analysis. *Assistance Systems*, such as smart tools and tablets, are connected to sensors. They serve as an interface for workers

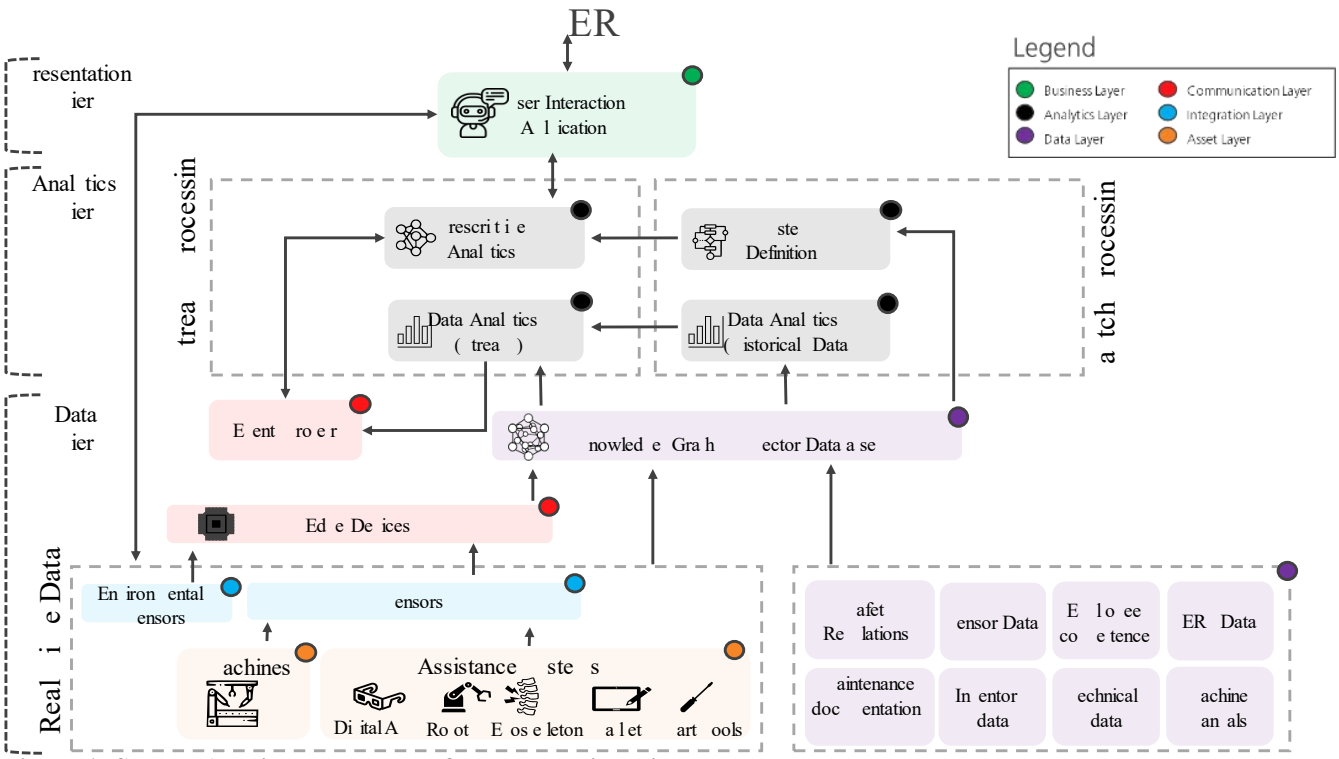


Figure 1: System Architecture Layout for empathetic assistance systems

on the ground, providing them with real-time guidance deri ed f ro t he s s te 's anal sis.

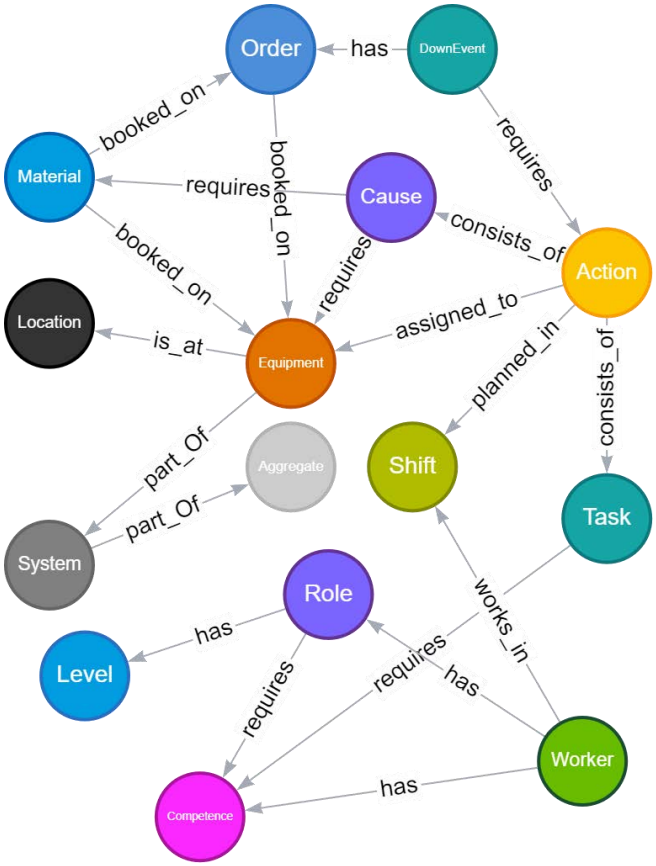


Figure 2: Data model of the KG based extended from (Kohl und Ansari 2023a)

Analytics Tier: The *Prescriptive Analytics* component is directly linked to the User Interaction Application. It processes the user's input, utilizing the Rasa conversational AI framework (Introduction to Rasa Open Source & Rasa Pro 2024), to generate actionable advice or maintenance recommendations. It uses advanced algorithms such as anomaly detection to suggest specific actions based on the analyzed data. The *System Definition* , powered by the Llama-2-70b-model (Touvron et al. 2023), functions as a reasoning framework that defines and orchestrates data analytic processes. It incorporates a multi-agent layer structure to process user input and determine the most appropriate action to take (Jiang et al. 2023). Therefore, necessary parameters and fitting data sources are determined to resemble the scope for aspired data analytics. The *Data Analytics (Historical Data)* component uses batch processing to analyze historical data to identify trends, patterns and potential issues based on past events. In contrast to the historical data analytics, the *Data Analytics (Stream)* component processes simulated real-time sensor data to offer immediate insights and detect current or impending issues, which is essential for real-time decision-making and alerts. The Data Analytics component utilizes multiple regression to forecast outcomes and incorporates K-means clustering to discover trends in historical data, as well as an Isolation Forest algorithm for anomaly detection. The foundational understanding of the Data Analytics (Historical Data) additionally augments the predictive real-time models to ensure a maximum of information for analysis.

Presentation Tier: The *User Interaction Application* component serves as the interface between the end-user and the chatbot system. It is where users interact with the chatbot, inputting queries and receiving responses. In this context, a simplistic User Interface featuring a Chatbot window was implemented, as illustrated in Figure 4.

Each tier in this architecture is intricately connected, allowing data to flow from the machines up through the system to enable real-time and predictive maintenance decision-making. The architecture is designed to maximize efficiency, reduce mean time to repair, and provide actionable insights through a user-friendly interface.

3.2. Modular Chatbot Layout

This chatbot layout is aligned with existing frameworks for developing multi-agent dialogue systems (Engelmann et al. 2023; Xi et al. 2023). It is depicted in Figure 3 and features a central User Agent linked to three specialized agents (Scheduling, Competency, Analyzer), all interfacing with an

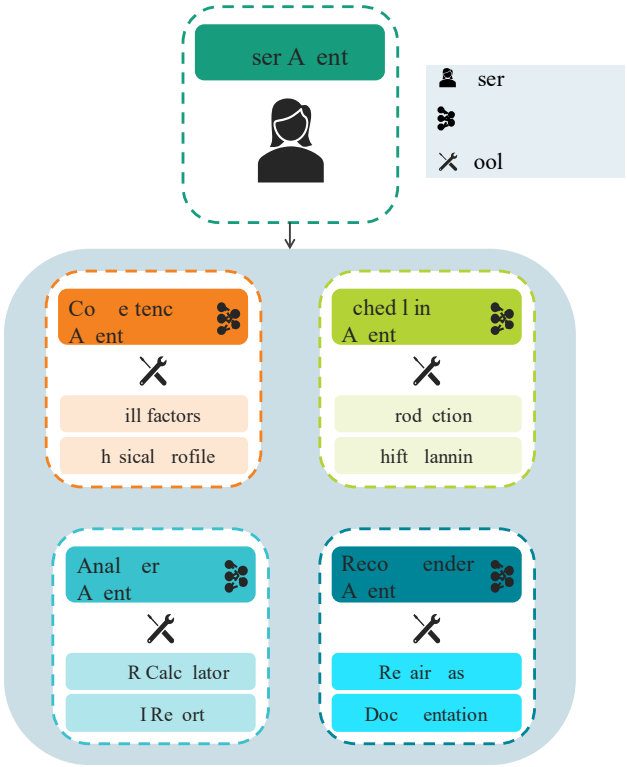


Figure 3: Interconnected Agent Layout for a modular maintenance chatbot architecture

LLM acting as a classification engine to determine which agent is triggered for a certain query. This User Agent represents the System Definition within the System Architecture, see Figure 1 and therefore determines which specialized agent is triggered subsequently. These specialized agents comprise of several tools and determine the correct tool usage for task-specific challenges. Moreover, the agents can interface with each other if the task requires agent collaboration. The proposed modular design allows for seamless integration of further agents and tools within agents to encounter novel challenges over time.

- **User Agent:** Channels user inputs to the appropriate specialized agents and consolidates their outputs for user communication. It is the link between Presentation and Analytics Tier.
- **Scheduling Agent:** Selects the production planning and shift planning tools based on user agent task instructions and operational needs. The optimization algorithm leverages provided data sources within the Data Tier and interacts with the Competency and Analyzer Agent to ensure fairness and efficiency while allocating shifts and schedule production.
- **Competency Agent:** Decides whether to analyze skill factors or physical profiles, aligning workforce tasks with individual skills and physical capabilities for optimal job assignment. Through its empathetic capability it continuously checks for physical and ethical alignment of worker tasks.
- **Analyzer Agent**: Chooses between MTTR calculation and KPI report analysis tools to assess maintenance effectiveness and identify areas for operational improvement. It provides recommendations such as prioritization or suggestions for automations.
- **Recommender Agent:** Has access to both historical and real-time data. When an anomaly is detected, it becomes operational. It offers similar historical failures, spare parts, and can store documentation in the KG.

Contrary to the User Agent the specialized agents interact with the Data Tier and leverage aforementioned RAG pipelines with KGs and vector databases to process dynamic and real-time information (Huang et al. 2024). This layout serves as an illustration of how agents can be utilized to allow dynamic maintenance strategies. The system architecture, see Figure 1, provides a high-level reference structure for the integration of new agents, such as a failure mode and effects analysis (FMEA) agent using the cause entity from the KG.

The architecture of this modular system integrates prompts as follows: The overarching system prompt guides the Chatbot, setting its function within a maintenance environment. This structure includes more specific prompts at subordinate levels. The User-Agent prompt functions analogously to a supervisory agent, tasked with identifying the most appropriate agent response to a user query. Each specialized agent operates under its own prompt; for example, the Analyzer Agent is responsible for generating reports based on historical or real-time data. This necessitates determining whether to initiate tools such as MTTR or KPI reports. Subsequently, this agent classifies the tool required for the task, parsing input parameters – such as the specific machine and time span – from the LLM. These parameters, where descriptions are also provided, are then employed within Python functions, with the resulting outputs fed back to the LLM, which then crafts responses based on these function outputs.

3.3. Regulative Considerations

In the context of implementing a maintenance chatbot, aligning with the EU AI Act's transparency obligations is essential for fostering trust and ensuring responsible use. The EU AI Act mandates that users are explicitly informed when they are interacting with AI systems, like chatbots. This requirement is critical in maintenance environments, where decisions can impact operational safety and efficiency. By disclosing the chatbot's AI nature, users are empowered to make informed choices about their engagement, understanding that they are consulting a machine for assistance. This transparency not only builds trust in the technology's capabilities and limitations but also reinforces the importance of human oversight in decision-making processes. Ensuring users are aware they are interacting with

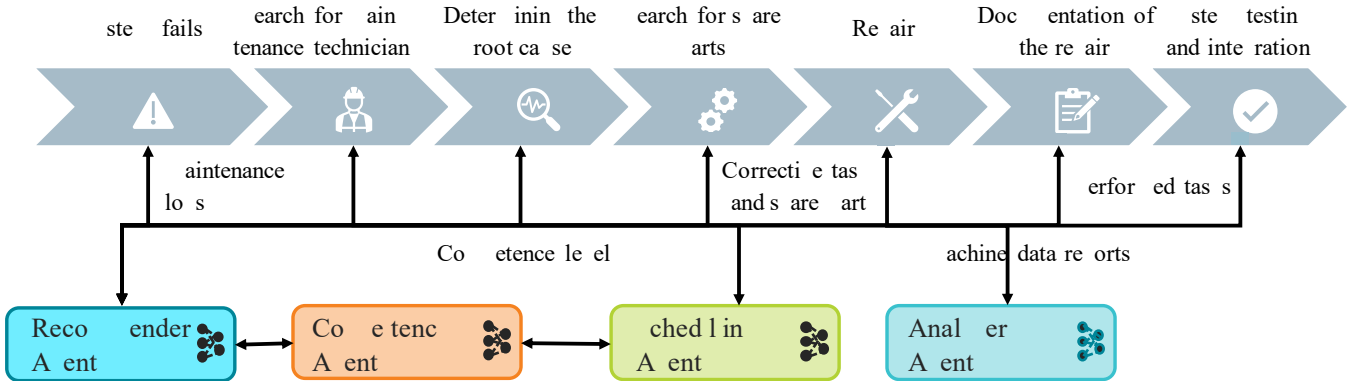


Figure 4: Maintenance process and its triggered agents as well as information flow

an AI helps maintain a balance between leveraging technological advancements and preserving human judgement and accountability in maintenance operations.

4. USE-CASE

The use case discusses a maintenance workflow in the railway industry, utilizing a chatbot for maintaining a cooling system., see Figure 4. The used cooling system provides sensor data about machine states as well as extensive manufacturer information. As it is one of the most frequently installed systems in Vienna's tramways and subways, extensive data on maintenance incidents in form of logs and spare parts is available. The data tier and the used maintenance data set consists of text-based, tabular industrial maintenance logs (Ansari 2020). The dataset was transformed in order to fit the structure of the Sequential QA (SQA) format by Microsoft (Iyyer et al. 2016) in order the be ideally processed by LLMs. In the use case scenario, the KG is constructed from maintenance logs exported from an ERP system, detailing machine failures and corrective actions, also integrates information on the equipment and spare parts used for repairs. The association of actions with required competence and the frequency of these actions by maintenance technicians serve to depict their competence levels (Ansari et al. 2023). Further, a vector database houses segments, specifically text excerpts, from work instructions and machinery documentation. For real-time data, the system monitors the current production schedule along with a simulated data stream of sensor readings from the machines. Additional stress levels of maintenance technicians are recorded for evaluation purpose.

Figure 5: Maintenance of a railway cooling system using a chatbot

4.1. Application of the Chatbot

The chatbot's supportive capability is discussed based on the standard end-to-end maintenance process see Figure 4. It consists of equipment failure, search for maintainer, identification of fault cause, search for spare parts, repair action, documentation of the maintenance process, reintegration of the machine. The following shows points of human interaction as well as autonomous chatbot within this process.

1. **Equipment failure**: The recommender agent is activated by an error notification, triggered by an anomaly in the real time data flow of the machine. Based on the error notification similar historical failures and corresponding actions are determined by semantic search of the task recommendation agent (Ansari et al. 2021).
2. **Search for maintainer**: The scheduling agent, competency agent and task recommendation agent exchange information about the production schedule, available maintenance personnel, their corresponding competencies, and the necessary tasks for failure resolution. According to that an allocation of the most fitting maintainer for the task is deducted.
3. **Identification of fault cause**: This stage marks the initial interaction between the maintainer and the chatbot. Utilizing the chatbot's knowledge, sourced from documents within the vector database, it can pose inquiries related to specific domains or machinery. Throughout this process, the human evaluates the tasks recommended by the chatbot for accuracy and corroborates them based on personal experience and the information furnished by the chatbot.
4. **Search for spare parts**: Once the tasks required for resolving the failure are identified, the task recommendation agent traverses through historical data in the KG to propose necessary spare parts.
5. **Repair action**: During the physical repair, the chatbot acts as an accessible source of pertinent information, offering guidance through machine documents or other necessary data from the vector database. Additionally, it can process requests for more detailed machine information, which are then thoroughly examined by the analyzer agent, e.g. asking for the mean time of repair.
6. **Documentation of maintenance process**: Building on prior interactions, the chatbot can autonomously create new connections within the KG and carry out the documentation process upon request from the maintenance personnel.
7. **Reintegration of the machine**: Finally, the chatbot guides through standard tasks to reintegrate the machine levering information from diverse work instructions.

The proposed integration of a chatbot within the standard end-to-end maintenance process, see Figure 4, represents a significant advancement in operational efficiency and precision. By embedding intelligent, autonomous capabilities at critical junctures of the maintenance workflow, from initial equipment failure detection to the reintegration of repaired

machinery, this model showcases a transformative shift towards more resilient and adaptive maintenance operations. The synergy between human expertise and artificial intelligence not only enhances the decision-making process but also optimizes resource allocation, reduces downtime, and enables empathic human-machine collaboration (Sorin et al. 2023).

4.2. Example: Analyzer Agent

To illustrate a potential maintenance workflow, we present an example of a chat where a user activates the Analyzer Agent. The dialogue demonstrates the triggered process chains, including the reasoning (highlighted in italics) and the tools employed. Due to space constraints, a detailed description of the various prompts is largely omitted. The chat ot is referred to as "AI." se rs can also follow the thought process through a collapsible dropdown segment in the interface, see Figure 6.

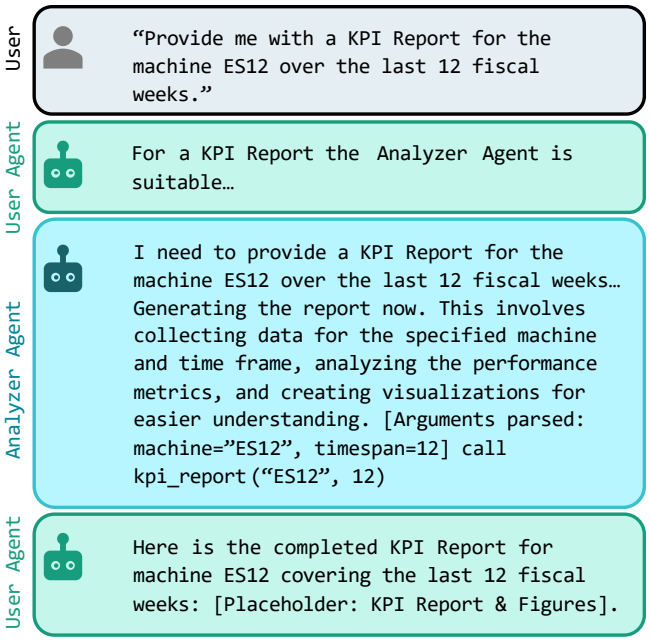


Figure 6: Example prompt of the Analyzer Agent

This example highlights how the chatbot, in its current state, bases its decisions, extending to the collaboration of multiple agents to optimize outcomes, such as in production planning.

4.3. Evaluation of the maintenance-chatbot

The evaluation is based on two types of maintenance tasks performed in Wien's pilot factory: a simple task for changing and cleaning a filter and a more complex task for changing the rotor, where the root cause is not clear. The depicted tasks require different competence levels in different areas. In the test scenarios, the maintenance chatbot demonstrated promising results for guiding the maintenance technicians through the root cause identification and for offering more detailed answers when needed, thereby reducing MTTR by 25% in comparison to the control group. In the case of the more challenging problem setting, the possibility of creating KPIs in natural language for deeper analysis reduced the MTTR by approximately 30% leading to an even higher impact. Furthermore, the dialogues are tailored to the individual competence levels, which permit queries and elucidations of (partial) steps, diminish the stress level and cognitive load, and facilitate a more empathetic conversational style. In addition, a ground truth dataset was constructed. Based on the logs, the appropriate agent or tool was triggered, and that the response was satisfactory in 83% of all cases. Specific tools, such as LangSmith (Ito et al. 2020), are currently under evaluation concerning integration for even better agent handling. Given the implementation of the LLaMA2 model within this chatbot, the Do-Not-Answer dataset (Wang et al. 2023) establishes a framework for safeguarding LLMs against potential risks. The efficacy of this dataset in mitigating harms will be further assessed in forthcoming studies through an adapted version tailored to evaluate the specific vulnerabilities and challenges posed by this chatbot.

5. Conclusion and Outlook

In summary, this investigation highlights a maintenance chatbot's significant efficiency over traditional systems in minimizing Mean Time to Repair (MTTR), thereby boosting operational efficiency and equipment effectiveness in manufacturing. Traditional NLP based systems show an improvement in MTTR of at least 20% in production environments, which is confirmed by preliminary investigations by (Ansari et al. 2023). Additionally independent studies on LLM based assistance systems (Noy und Zhang 2023) show even higher potentials in operational efficiency, well in line with the first tests of the detailed maintenance chatbot in the pilot factory use cases. Leveraging advanced NLP and machine learning, the chatbot surpasses conventional systems by integrating ERP data and identifying relationships for enhanced maintenance insights, significantly reducing cognitive load and stress.

Looking ahead, the scalability and generalizability of the maintenance chatbot are poised for improvement with the multi-agent systems, and causal AI. AutoGen frameworks (Wu et al. 2023) are anticipated to refine the chatbot's content generation and adaptation capabilities, enabling reciprocal learning. Multi-agent systems promise to distribute problem-solving tasks effectively, improving maintenance operations' efficiency. Meanwhile, causal AI could provide a deeper understanding of the complex causal relationships within maintenance data systems, offering more accurate step-by-step solutions.

Future directions indicate that maintenance chatbots could overcome current limitations and adapt across various manufacturing settings. This flexibility is key to meeting the

sector's varied needs, marking a significant advancement in CAS for maintenance. Driven by improvements in data integration, natural language processing, and causality understanding, this represents a crucial step in manufacturing's digital transformation.

ACKNOWLEDGEMENT

The authors would like to acknowledge the financial support of this research work by the Fraunhofer-Gesellschaft as part of the Fraunhofer flagship project EMOTION.

REFERENCES

A Rasheed, asa n; A d ls ala , o ha a d ssa ; We er , Christian; Fathi, adj id (2024): ortin t d ent Decisions on earnin Reco en dations: An a sed Chat ot with no wled e Gr a h Co nte t ali ation for Con er sational E lai na ili t and e ntorin . O nline erfü ar nter htt :// ar i .or / df/ 2401.08517 3.

A arwal, Oshin; Ge, e i n ; ha er i, ia a ; Al Rfo , Ra i (20 20): no wled e Gr a h ase d ntheti c Cor s Generation for no wled e Enhanced an a e odel r e trainin . O nline erfü a r n ter htt :/ /ar i .or / df/2 010.12688 2.

Ala i, a r a ; ei dner, Doroth E .; o s a i, Re a (2024): n owled e ana e ent e rs ecti e of Generati e Artificial Intelli en ce. In: *JAIS* 25 (1), . 1–12. DOI: 10.17705/1jais.00859.

Ansari, Fa el (2019): nowled e ana e e nt 4.0: he oretical and ract ical Considerations in C er h sic al r od ctio n st e s. In: *IFAC-PapersOnLine* 52 (13), . 1597–1602. DOI: 10.1016/j.ifacol.2019.11.428.

Ansari, Fa el (2020): Cost a sed te t n derstandin to i ro e ai ntenance n owled e intelli en ce in an f act r in enter r ises. In: *Computers & Industrial Engineering* 141, . 106319. DOI: 10.1016/j.cie.2020.106319.

Ansari, Fa el ; Glawar, Ro e rt; Ne eth , anja (2019): r i a: a resc ri ti e ainte nance o del for c er h s ical ro d ction s s te s. In: *International Journal of Computer Integrated Manufacturing* 32 (4 5), . 482–503. DOI: 10.1080/0951192X.2019.1571236.

Ansari, Fa el ; o ld, h ili ; ho r eh, a rjan (2020): A nowle d e as ed a r oach for re resenti n j o holde r rof ile toward o ti al h a n– ac hine colla orati on in c er h sical ro d ction s s te s. In: *CIRP Journal of Manufacturing Science and Technology* 28, . 87–106. DOI: 10.1016/j.cir j. 2019.11.005.

Ansari, Fa el ; ohl, in s ; Giner, Ja o ; eier, orst (2021): e t inin f or AI enhanced fail re detection and a ail a ili t o t i i at ion in r od ctio n s s te s. In: *CIRP Annals* 70 (1), . 373 –376. DOI: 10.1016/j.cir .202 1.04.045.

Ansari, Fa el ; ohl, in s ; i hn, Wilfried (2023): A co etence a sed lan nin ethodolo f or o ti i i n h an reso rc e allocation in ind s trial ai ntenance. In: *CIRP Annals* 72 (1), . 3 89–392. DOI: 10.1016/j.cir .202 3.04.050.

e sin er, hil i ; ejnos a, Daniel; Ansari, Fa el (2024): Res onsi le AI (RAI) in a n fact r in : A Q alitati e Fra e wor . In: *Procedia Computer Science* 232, . 813–822. DOI: 10.1016/j. r ocs.2024.01.081.

i rhane, A e a; asir a deh, Atoosa; esli e, Da id ; Wachter, an dra (2023): cie nce in the a e of lar e lan a e odels. In: *Nat Rev Phys* 5 (5), . 277–280. DOI: 10.1038/s42254 023 00581 4.

rown , o .; a nn, e nja in; R der , Nic ; iah , e lanie; a lan, Jared; Dhariwal, raf lla e t al. (2020): an a e odels are Few hot ea rners. Online erfü ar n ter htt :// ar i .or / df/2 005.14165 4.

r räf, ete r; Danna fel, a tthias; Adlon, o ia s; Föhlisch, Nils (2021): Ada ti e asse l s st e s for ena lin a ile asse l – E irical anal s is foc s in on co niti e wor e r assistance. In: *Procedia CIRP* 97, . 3 19–324. DOI: 10.1016/j. r ocir.2020.05.244.

Elo n do , n a; ann in , a ; ish in, a ela ; Roc , Daniel (2023): G s are G s: An Earl o o at the a or a r et I act o tential of ar e an a e odel s. Online erfü ar nter htt :// ar i .or / df/ 2303.10130 5.

En el an n, Dé ora C.; anis son, Alison R.; ieira, Renata; ü ner, Jo i Fred; asca rdi, i iana; ordini, Rafael . (2023): A ID — A Fra e wor f or the De el o e nt of lt i A ent Intentional Dialo e s te s. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*, . 1209–1217.

E ro ean Co issi on (2023): E lo ent and social de elo ents in E ro e 202 3. e o r : li cations Office of the E ro ean nion (E lo ent and social de elo ents in E ro e 202 3).

E ro ean Co issi on (2024): E AI Act. Article 52, rans a renc O li atio ns for ro iders and ser s of Certain AI s te s and G AI odel s. Online erfü a r n ter htt s ://www.e aiact .co / article/52, let t e rüf t a 27.03.2024.

Fensel, Dieter; Şi şe , tcan; An ele, e in; a a n, Elwin; är le, Elias; an asi , Ole s andra et al. (2020): Introd ctio n: What Is a n owled e Gra h? In: Dieter Fensel, tc an Şi şe , e in An el e, Elwin a a n, Elias ärle, Ole sa ndra anasi et al. (.): nowled e Gra hs. Cha : rin er I nternational li shin , . 1–10.

Freire, a el erna n; a nic er , ar ath r endranadha; R i Arenas, antia o ; R s á , Z oltán; Niforatos, E a n elos (2023): A Co niti e Assistant for O er ators: AI owered no wled e harin on Co l e s te s. In: *IEEE Pervasive Comput.* 22 (1), . 50–58. DOI: 10.1109/ R .2022.3218600.

Go al o ri e la, Ro erto; Garrido e rchan, Ed ar do C. (2023): ChatG is not all o need . A tat e of the Art Re iew of lar e Generati e AI odels. Online er fü ar n ter htt :// ar i .or / df/2 301.04655 1.

an, Y ; ao, Jin wen (2024): Re ol tioni in har a: n eil in t he AI and rends in the har ace tica l Ind s tr . Online erfü a r n ter htt :/ /ar i .or / df/2 401.10273 2.

an , X ; i , Weiwen; Chen, Xiaolon ; Wan , Xin e i; Wan , ao ; ia n, Def et al . (2024): n derstandin the lan nin of a e nts: A s r e . Online e rfü a r n ter htt :/ /ar i .or / df/2 402.02716 1.

Introd ctio n to Rasa O e n o rce Rasa ro (2 024). Online erfü ar n ter htt s: //rasa.co /docs/rasa/, let t a t ali siert a 22.03.2024.

Ito, a i; r i a ashi, ats i; ida a, a satoshi; i, J n ; In i, en taro (2020): a n s it h: An Interacti e Acade ic e t Re isio n s te . Online erfü ar nter htt :// ar i .or / df/ 2010.04332 1.

I er , o hit; Yih, Wen ta ; Chan , i n Wei (2016): Answerin Co li cated Q e stion Intents E resse d in Deco osed Q estion e q ences. Online er fü ar nter htt :/ /ar i .or / df/1 611.01242 1.

Jian , Z hiqi ; Rashi , ash r r ; anc hal, n jal; Jasi , a h ood; a r ha d, Ali; Riahi, ari et al. (2023): Co nit ot s: Creatin and E al ati n A lt i A en t Chat ot lat for for lic In t Elicitation. In: *Proc. ACM Hum.-Comput. Interact.* 7 (C CW 1), . 1 –32. DOI: 10.1145/3579469.

Jin , Z hi; , Yon e; an, Yi n; Y an, o; X , ai n; i , Ch njian et al. (2024): When ar e an a e odels e et ector Data ases: A r e . Online erfü ar nte r htt :/ /ar i .or / df/2 402.01763 2.

an , Y e; Cai , Zhao; an, Chee Wee; a n , Qia n; i , ef (2020): Nat ral lan a e r ocessin (N) in ana e e nt research: A literat r e re iew. In: *Journal of Management Analytics* 7 (2), . 139 –172. DOI: 10.1080/23270012.2020.1756939.

erna n Freire, a el; Foosherian, ina; Wan , C haofan; Niforatos, E a n elos (2023): arnessi n ar e a n a e od els for Co niti e Assistants in Factories. In: i nha e e, Cos in ntean , arti n orcheron, Johanne ri as nd ara h heres ö l el (.): r oceedin s of the 5th International Conference on Con e rsational ser I nterfaces. C I '23: AC co nference on Con ersat ional ser Interfaces. Eindho en Netherlands, 19 07 2023 21 07 2023. New Yor , NY , A: AC , . 1–6.

ohl, in s ; Ansari, Fa el (2023a): A no wled e Gra h ase d ea rnin Assistance ste s for I nd st rial a intenance, in r ess.

ohl, in s ; Ansari, Fa el (2023) : Chat ots in der Instandhalt n s lan n : Ind st rielle Anwend n s fälle nd ü nfti e ers e ti en: Ö IA on r ess.

o stolani, Da i d; Wollendorfer, ic hael; c hl nd, e astia n (2022): Er o a s : owards Inter ret a le a nd Accessi le A t o ated Er o no ic Anal sis . In: 2022 IEEE 3rd International Conference on a n a chine ste s (IC) . 2022 IEEE 3rd International Conference on an a chine s te s (IC). Orlando, F , A, 17.11.2022 19.11.2022: IEEE, . 1–7.

i, Y nqin ; R a an, hi a ar; Co hen, a l; t arl , i nil (2021): Desi n of nowled e Gra h in a n fact r in er ices Disc o er . In: ol e 2: an fac t rin rocesse s; a n fact rin ste s; Na no/ icro/ eso a n fact rin ; Q alit and Relia ili t . A E 2021 16th International a n fact rin ci ence and En ineerin Co nference. irt al , Online, 21.06.2021 25.06.2021: A erican ociet of ec hanical En in eers.

ist l, Fran Ge or ; Fisc her, Jan; We rich, i chael (2021): owards a i lati on ase d Con er sational Assistant for the O e ration and En ineerin of ro d ction la nts. In: 2021 26th IEEE International Conference on E er in

echnolo ies a nd Factor A to ation (E F A). 2021 IEEE 26th International Conference on E er in echnolo ies and Factor A to ati on (E FA). asteras, we den, 07.09.2021 10.09.2021: IEEE, . 1–4.

a r a ria, i iana; c hiewec , Ale ander (2019): he Di ital h read in Ind s tr 4.0 . In: Wolf a n Ahrendt nd il ia i e th a ia arifa (.): I nte rated For al et hods, d. 11918. Cha : rin er I nternational li shin (ect re Notes in Co ter cience), . 3–24.

a r , ene di t G.; Ra ch, E rwin; at t, Do ini . (2021): Wor er assi stance s s te s in an f act r in : A re iew of the state of the art and f t r e directions. In: *Journal of Manufacturing Systems* 59, . 228–250. DOI: 10.1016/j.j s .2021.02.017.

No , ha e d; Zhan , Whitne (20 23): E eri ental e ide nce on the ro d ct i it ef fects of enerati e artificial intelli en ce. In: *Science (New York, N.Y.)* 381 (6654), . 187–192. DOI: 10.1126/science.adh2586.

OECD Artificial Intelli ence a e rs (2024).

an, hir i; o, in hao; Wan , Y fei ; Chen, Chen; Wan , Jia ; W , Xindon (20 24): n if in ar e an a e od els and nowled e G ra hs: A Road a . In: *IEEE Trans. Knowl. Data Eng.*, . 1–20. DOI: 10.1109/ DE.2024.3352100.

o orn i, asti an; Constantinesc , Ca r e n (2021): Desi n and Confi rat ion of Di ital Assistance st e s in a n al Asse l o f ariant rich r od cts ased on C s to e r Jo r ne a i n . In: *Procedia CIRP* 104, . 1777–1782. DOI: 10.1016/j. ro cir.2021.11.299.

Ra o , o ss a ; i , Xi n an; ern, Ro a n (2023): I ro in F EA Co r ehensi il it ia C o on ens e no wled e Gr a h Co letion echniq es. In: *IEEE Access* 11, . 1 27974–127986. DOI: 10.1109/ACCE .202 3.3331585.

DIN 91345, 2016: Referen a rchite t r odell Ind s trie 4.0 (RA I 4.0).

Ro er o, Da i d; tah re, Johan (2021): owards he Resilient O e rator 5.0: he F t re of Wor in art Resilient an fact r in s te s. In: *Procedia CIRP* 104, . 108 9–1094. DOI: 10.1016/j. ro cir.2021.11.183.

Rožanec, Jože .; , Jin hi; R n i , Jan; Š r janc, aj a; l adenić, D nja ; Fort n a, l až et al. (2022): Actiona le co niti e twins for decision a in in a n f act r in . I n: *International Journal of Production Research* 60 (2), . 452–478. DOI: 10.1080/00207543.2021.2002967.

a oo, .; he h awat, D. (2024): Enhancin r edicti e a intenance in an Oil Gas Refiner sin Io , AI : An Generati e AI ol tion. In: Da 3 Wed, Fe r ar 14, 2024. International etr ole echnolo C onference. Dhahran, a d i Ara ia, 12.02.2024 12.02.2024: I C.

h in, Won; a n, Jeon n; Rhee, Wonjon (202 1): AI assistance for r edicti e ai ntenance of renewa le e ner s ste s . In: *Energy* 221, . 119775. DOI: 10.1016/j.ener .2021.119775.

o rin, era; rin, Danna; ar ash, Yiftach; onen, Eli; Charne , Ale ander; Nad arni, Girish; la n , E al (2 023): ar e a n a e o dels (s) and E a th – A ste ati c Re iew .

n, Yichen ; Zhan , Qi; ao, Jinson ; , Y qia n; i , h i i n (2024): E ower in di ital twins with lar e lan a e odels for lo al t e oral feat re learnin . I n: *Journal of Manufacturing Systems* 74, . 83–99. DOI: 10.1016/j.j s .2024.02.015.

o ro n, o; arti n, o is; ton e, e i n; Al er t, ete r; Al ahai ri, A j ad; a aei , Yas ine et al . (2023): la a 2: O en Fo ndati on and Fine ned Chat odels. Online erfü ar nter htt :// ar i .or / df/ 2307.09288 2.

Wan , Y ia ; i , aonan; an, X don ; Na o , r esla ; a ldwin, i oth (2023): Do Not Answer: A Dataset for E al ati n af e ards in s. Online e rfü ar nte r htt s:/ /doi.or / 10.48550/arXi .2308.13387.

Wan , Z h o; ai, Xiaolian ; Zhan , h s hen ; i llin h rst, ar ; e, Wei in ; Wan , e n et al . (2022): A co r ehensi e re ie w of a ente d realit ased instr ction in an al asse l , trainin a nd re air . In: *Robotics and Computer-Integrated Manufacturing* 78, . 102407. DOI: 10.1016/j.rci . 2022.102407.

W , Qi n n; a nsal, Ga a n; Zhan , J ie ; W , Yiran; i, e i in; Zh , E r an et al . (2023): A toGen: Ena lin Ne t Gen A li cations ia lti A en t Con er sation. Online erfü ar n ter htt :// ar i .or / df/2308. 08155 2.

Xi, Zhihen ; C hen, Wen ian ; G o, Xi n; e, Wei; Din , Yiwen; on , o an et al . (2023): he Rise and otential of ar e an a e o del a sed A en ts: A r e . Online erfü ar nter htt :// ar i .or / df/ 2309.07864 3.

Y , on Qin (2021): D na ic C a sali t no wled e Gra h Gene ration for ort in the Chat ot ealthcare

ste . In: ohei Arai, ri a a oor n d Rah l hati a (.): r oceedin s of the F t re echnolo ies Conference (F C) 2020, ol e 3, d. 1290. Cha : ri n er International lishin (A d ances i n Intelli ent st e s and Co ti n) , . 30–45.

Y , Wenhao (2022): Retrie al a ented Generation across eter o eneo s nowled e. I n: Da hne I oli to, i n ian arold i, ar ia eonor ac heco, Danqi Chen nd Nianwen X e (.) : roceedin s of the 2022 Conference of the North A e rican Cha ter of the Association for Co tational in istics: an a n a e echnolo ies: t d ent Research Wor s ho . r oceedin s of the 2022 Conference of the North A e rican Cha ter of the Association for Co ta tional i n isti cs: a n an a e echnolo ies: t dent Research Wor sho . r id: eatt le, Washin to n + Online. tr o ds r , A, A: Association for Co t ational in istics, . 52–58.

Zhao, Andrew; an , Daniel; X , Q entin; in, a tthie ; i , Yon Jin; an , Ga o (2023): E e : A e nts Are E e riential earners. Online er fü ar nter htt :/ /ar i .or / df/2 308.10144 2.

Zhen , Xia ochen; , Jin h i; iritsi s, Di i tris (2022): he e er en ce of co niti e di it al twin: isio n, challen es and o ort n ities. In: *International Journal of Production Research* 60 (24), . 7 610–7632. DOI: 10.1080/00207543.2021.2014591.

Zho , in; i, Xin ; i , ian a n; X , ai ho ; i , Wei; ao, Jinson (2024): Ca s al G : Ind s trial str ct re ca sal nowle d e enhanced lar e lan a e o del for ca se anal sis of q alit ro le s i n aeros ace r od ct an f act r in . In: *Advanced Engineering Informatics* 59, . 102333. DOI: 10.1016/j.aei.2023.102333.

Zh , Yin ha o; Ren, Chan ; Xie, hi n; i , h ai ; Ji, an an; Wan , Zi ian et al. (2024): REA : R AG Dri en Enha nce ent of lt i o dal Electronic ealth Records Anal sis ia a r e an a e odels. Online erfü ar nter htt :// ar i .or / df/ 2402.07016 1.

Zi art, anja; chl n d, e a stian (2020): E a l ati on of A e nted Realit echnolo ies in an f act r in – A it erat r e Re i ew. In: Isa el . N nes (.): Ad ances in an Factors and ste s Interaction, d. 1207. Cha : rin er International lishin (Ad a nces in Intelli en t ste s and Co tin), . 75 –82.

BIOGRAPHIES

L. Kohl, Dipl.-Ing., has been a research assistant at the Institute of Management Sciences at TU Wien and at Fraunhofer Austria Research GmbH in factory planning and production management since September 2019. At Fraunhofer Austria, Linus Kohl leads the group for production optimization and maintenance management. Mr. Kohl studied industrial engineering - mechanical engineering at TU Wien. His work focuses on maintenance - the data-driven analysis and optimization of machines and systems using AI-based assistance systems. Linus Kohl is largely responsible for establishing the areas of retrofitting, maintenance as a service and industrial cognitive system at Fraunhofer Austria.

P. Besinger joined Fraunhofer Austria Research GmbH in 2021 in the role of research assistant, working on the development, integration and operation of AI-based software at companies and the associated requirements engineering. He completed his aste r's de r ee in ind strial en ineerin - Machine Engineering at TU Wien in 2021. Mr. Besinger received three TU Wien merit-based scholarships from 2018 - 2021 for outstanding performance. He is particularly interested in the application as well as implementation of Responsible AI in an industrial context to strengthen trust in AI-models and minimize socially harmful impacts of AI-systems.

S. Eschenbacher S. Eschenbacher is currently pursuing a aster's de r ee in AI En i neerin at the n i er sit of Applied Sciences Technikum Wien. She joined Fraunhofer Austria GmbH in 2022. Her research interest lies in the application of LLM and multi agent frameworks and on how these technologies can be deployed in real-world production environments to drive innovation and support the digital transformation in the industry.

F. Ansari is full professor at TU Wien and chair of Production and Maintenance Management and at the same time he serves as the head of strategic projects at Fraunhofer Austria. He conducts interdisciplinary research at the intersection of AI, Industrial Engineering and Production Management, where maintenance plays a central role. His interdisciplinary background is underlined by a degree in mechatronics and a dissertation in computer science (summa cum laude) at the University of Siegen. With his international involvement in various scientific associations (IEEE, IFAC, IALF), as well as his habilitation in Industrial Engineering, entitled "Management of Knowledge Intelligence in Human-centered Cyber Physical Production Systems", Dr. Ansari has established his role as an important part of the international research community.

Leveraging Generative and Probabilistic Models for Diagnostics of Cyber-Physical Systems

Alvaro Piedrafita[1], Leonardo Barbini[2],

[1,2] *TNO-ESI, Eindhoven, The Netherlands*
alvaro.piedrafitapostigo@tno.nl
leonardo.barbini@tno.nl

ABSTRACT

A critical task for system operators is the precise identification of the root causes underlying an error situation. This identification is fundamental in deciding optimal maintenance actions, such as replacing a component versus calibrating it. However, the actual causes of an error are often neither measured nor unique. The measured quantities are the result of complex interactions between different error causes and system variables. Root cause identification in this context becomes a matter of inferring hidden causes from their measurable effects. This challenge is notably pronounced in cyber-physical systems comprising control loops. Control mechanisms, integral to maintaining system performance, introduce a layer of complexity in diagnostics and ultimately complicate the isolation of the underlying causes of errors. To address this challenge, we introduce a two-step approach to derive the hidden causes as a statistical inference task. First, we develop a generative model leveraging existing control software and expert-based insights into the mechanisms of errors, i.e., a simulator of synthetic data given some hidden error causes. Then, we transform the generative model into a probabilistic program on which statistical inference can be executed within a probabilistic programming language framework. This inference effectively estimates the hidden causes given some measured data from the system. Being intrinsically a statistical approach, these inferences come with a confidence interval. We applied this methodology to an industrial printer's sheet transport belt, operating in a closed-loop configuration. Our approach successfully discerned the contributions of three distinct hidden causes to the belt's deviation from its intended position. This paper highlights the efficacy of generative modeling followed by a probabilistic programming approach in unraveling complex interactions within cyber-physical systems for optimal maintenance.

1. INTRODUCTION

In order to match the increasing market demands on overall equipment effectiveness, industrial manufacturers of cyber-physical systems need efficient methods to diagnose system malfunctions. In practice, finding the root cause of such malfunctions is challenging for several reasons, ranging from technological to human and organizational ones.

On the technological side, the high demands on performance lead to increasingly complex systems with many intertwined control mechanisms that obscure the path from a root cause to its measurable effects (Borth & Barbini, 2019). The lack of direct observability for each cause of malfunctions forces the diagnostic to infer the many root causes from their effects on the few measured observables. Moreover, these observables are often not measured for diagnostic purposes but rather for control and performance ones, i.e. are indirect.

On the human and organizational side, the knowledge needed to solve difficult diagnostic cases is within the design and engineering departments, while the responsibility of offering diagnostic support lies on the service department (van Gerwen, Barbini, & Nägele, 2022). The transfer of the necessary knowledge is thus a difficult process that relies on expensive escalation-based approaches, i.e. design and engineering departments are called in by service to support the diagnostic reasoning. Finally, the struggle to timely train the service personnel capable of executing the needed diagnostic reasoning is a growing concern in the face of relentlessly increasing system complexity.

To tackle the points above we propose a method that focuses on two pillars. First, capture in models the knowledge of the system behavior, e.g. control loops, together with its failures. This should be done iteratively within the design and engineering departments, by incrementally incorporating knowledge on failures occurring in the field. Second, support the diagnostic reasoning process by performing statistical inference on the above models together with data from an error situation in the field, inferring the hidden causes of errors in

a Bayesian way. This is the contribution of this methodology to the service department.

The rationale behind the proposed approach is that humans have the knowledge and the inclination to reason *forward*, i.e. in a simulation-like manner from the causes of a failure towards the resulting effects. Many such simulation models are readily available in industrial companies. Conversely, it is more challenging for humans to perform *inverse* diagnostic reasoning from the effects towards the causes: they need computational support to do so. In this paper we leverage the available system expertise and modeling capabilities of humans, with statistical inference tools to achieve diagnostic reasoning support.

The remainder of the paper is organized as follows: below we give an overview of the relevant literature. In Section 3 we introduce the details of our approach. In Section 4 we first apply our methodology to synthetic data and then to real data from an industrial system, finally we conclude our paper and give directions for future research in Section 5.

2. Literature review

The proposed method has its foundations in model-based diagnostics (De Kleer & Kurien, 2003) and specifically in its probabilistic implementation with Bayesian networks (Lucas, 2001; Srinivas, 1995). In this context, Bayesian networks, a type of probabilistic graphical model, are used to infer the likelihood of causes based on observed data via Bayes' theorem. The quantity of interest for the diagnosis is the posterior probability of cause C given observations O, computed as $P(C|O) = P(O|C) \cdot P(C)/P(O)$.

The present paper extends the previous work in two directions. First, we model and reason with continuous random variables, rather than discrete. This is fundamental when tackling performance issues, i.e. scenarios where the system's components are not described by a neat dichotomy of states, such as *normal* or *abnormal*, but rather sit in a continuous spectrum of states. Second, we model and reason on dynamic processes rather than on static ones. This is needed when diagnosing systems with feedback control loops and when the cause of failures shows a time-dependent behavior. In the literature, such systems are often modeled with dynamic Bayesian networks (Bartram & Mahadevan, 2015), but this is cumbersome and very quickly results in very large models, so we propose a different approach.

In this paper, we perform statistical inference on dynamic models with continuous random variables by using a probabilistic programming paradigm (van de Meent, Paige, Yang, & Wood, 2018). The proposed probabilistic programming approach can be seen as a generalization of Bayesian filtering and smoothing methods (Särkkä & Svensson, 2023) such as Kalman filters and particle filters. Several methods have been introduced in the probabilistic programming literature to perform such statistical inference, sampling-based methods like Markov chain Monte Carlo (van de Meent et al., 2018), gradient based methods (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2017) and analytic methods like message passing (Cox, van de Laar, & de Vries, 2019), or combinations thereof (Cox et al., 2019). In this paper we rely on Markov chain Monte Carlo using the Python library Numpyro (Phan, Pradhan, & Jankowiak, 2019). The proposed methodology makes use of simulation models to generate synthetic data for validation and fine-tuning of the inference models. The use of synthetic data has been explored before in other fields, see (Tremblay et al., 2018) on the use of synthetic data in deep learning, and (Cranmer, Brehmer, & Louppe, 2020) for a discussion on the use of simulation for inference.

3. Methodology

Our methodology is schematically represented in Figure 1. It uses two models, *simulation* and *inference*, and develops in three phases, *creation, validation* and *usage* phases. These are represented with different colors in the figure; orange, blue and green, respectively.

In the *creation* phase, we first compile a *simulation* model using knowledge of the system, thus re-using already available control models, and augmenting these with (conjectured) models of failure mechanisms. The latter heavily relies on expert knowledge based on historical failures. The *simulation model* outputs synthetic data given a single or a combination of failure mechanisms. The *simulation model* is then transformed into an *inference model*. This transformation is not computational, i.e. it requires additional modeling. For example, some aspects of the simulation might be deemed irrelevant or negligible and dropped from the inference. Further, one could decide to decompose a single *simulation model* into multiple *inference models*. We will return to this in Section 4.2.

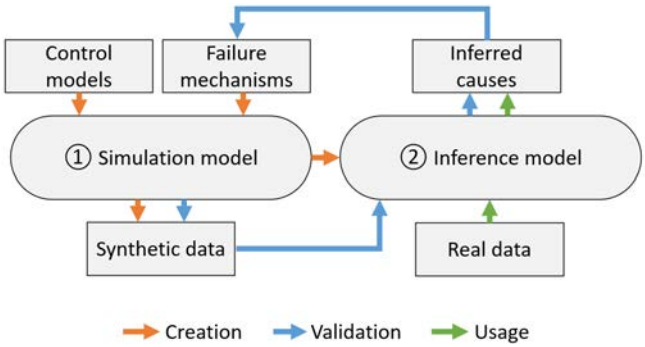


Figure 1. Schematic representation of the proposed methodology

In the *validation* phase, the goal is to verify that the constructed inference model indeed provides an inverse to the

data generation process. We do so by testing whether the model can correctly infer the hidden causes of failures in synthetic data generated in simulation models. We repeat this for different types and combinations of failures.

Finally, in the *usage* phase, we use the *inference model* on the real data coming from a system in the field to infer the hidden causes of failures. The optimal service action, e.g. part replacement versus cleaning, is then decided based on these inferred causes.

4. APPLICATION

We apply the proposed methodology to a subsystem of a Canon Production Printing (CPP) industrial printer. This subsystem contains a conveyor belt that rests horizontally on two cylinders. The cylinders rotate at a variable speed and transmit this movement to the belt. For this subsystem, it is required that the belt is at the center of both cylinders, perpendicular to the directions of its movement. To fulfill this requirement, one of the cylinders can be tilted by raising or lowering it. The mechanism responsible for this tilting is driven by a motor. In the remainder of this paper, we will refer to it as the Z-position motor. This tilting causes the belt to slide up or down the cylinder each revolution by an amount proportional to the Z-motor position. Every few revolutions the position of the belt is measured and a correction is computed by a Proportional Integral (PI) controller, resulting in an adjustment of the Z-motor position. This steering action is necessary to counter the various causes that make the belt drift away from its intended position.

Our goal here is to discern the unknown causes of this drift and to infer their strength, given the available data on the belt and motor positions over time. This is crucial from a maintenance perspective, to define the best service action in those cases in which, despite the control mechanism, the belt goes out of its intended position. Following our methodology, we first make a model relating the known, i.e. measured, and the unknown variables of this system. Then we conjecture the functional form of the unknown variables to create a complete simulation model. In the next Section, we use the equations of the PI-controller for the former and expert knowledge for the latter.

4.1. Simulation model

Every step of the PI-controller begins with a measurement of the belt position. This belt position must be a function of the previous belt position, the previous motor correction, and the drift incurred between the current measurement and the previous one. Based on the current positions of both the belt and Z-motor, the position of the latter is updated by a PI controller with the goal of returning the belt to its intended position. The equations modeling this behavior are:

$$\begin{cases} \mathsf{belt}_k = & \mathsf{belt}_{k-1} - \alpha \cdot \mathsf{motor}_{k-1} + \mathsf{drift}_k \\ \mathsf{integral}_k = & c_{int}(\mathsf{belt}_k + \mathsf{belt}_{k-1}) + \mathsf{integral}_{k-1} \\ \mathsf{motor}_k = & c_{prop}\mathsf{belt}_k + \mathsf{integral}_k \end{cases} \quad (1)$$

Where α, c_{int}, and c_{prop} are known proportionality constants and subscripts $(\cdot)_k$ corresponds to the value at sample k. All three quantities are measured. Notice that in Equation (1) the last 2 equations are taken directly from the implementation of the controller.

Not contained in these equations is the condition that the steering motor stays within a bounded range. If the necessary correction is outside these bounds, the motor will stay at the limit of its range, causing the belt to drift outside of its desired position. Throughout this paper, we assume that the motor and the belt position sensor never fail. This assumption can be relaxed, if necessary, and the proposed methodology can still be applied.

In Equation (1) the drift can be computed at all times since it is a function of the belt and motor positions, both measured. What remains unknown are the different error mechanisms and how they add up to the total drift. For this, we use expert knowledge.

We conjecture that the drift results from the linear combination of five causes:

- Calibration: the belt might not be completely horizontal when the Z-motor is at position 0. This results in a constant calibration error c.
- Misalignment: the belt might not be well aligned with the previous component of the printer, which results in pages coming into the belt with a lateral velocity relative to the direction of motion of the belt, causing drag to one side. This results in a constant misalignment error m that is present only when the machine is printing.
- Degradation: the belt material might wear out and deform over time, resulting in a time-dependent drift D_k. We conjecture this degradation to be exponential and with an unknown deformation direction.
- Sheets: when the pages make contact with the belt, they might cause a perturbation to its position, depending on the properties of the pages. This would result in a train of pulses P_k with varying amplitude and width, present only when the machine is printing.
- Noise: we finally conjecture that all other sources of error add up to a Gaussian term $\varepsilon_k \sim \mathcal{N}(0, \sigma)$ with unknown variance and zero mean.

These causes are described by the following equations:

$$\mathsf{drift}_k = c + \mathsf{print}_k(m + P_k) + D_k + \varepsilon_k \quad (2)$$

$$D_k = s(4^{\delta k} - 1) \quad (3)$$

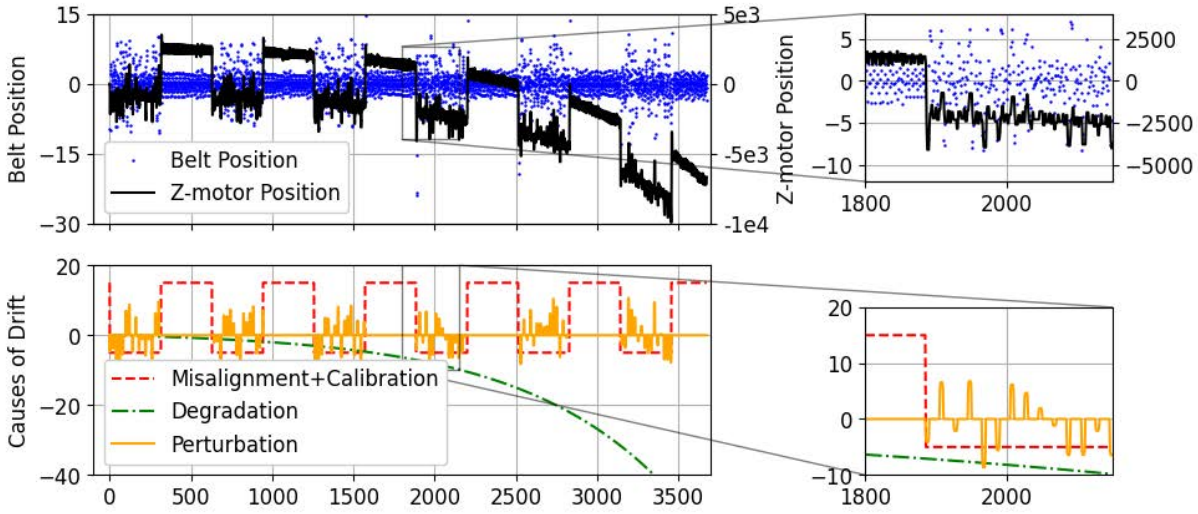


Figure 2. Example synthetic data produced with the simulation model. Observe how the motor displacement follows the sum of the three hidden contributors to the belt drift. See main text for a detailed explanation.

In equations (2,3) we can see the decomposition of the drift into its different terms, with the conjectured form of the degradation as being an exponential with exponent parameter $\delta \geq 0$ and sign $s \in \{-1, 1\}$. print_k is the variable that represents whether the machine is printing and takes values in $\{0, 1\}$. The variables c, m, D_k and P_k in Equation (2) are unknown, while print_k and drift_k are known quantities. In the interest of brevity, we have not included here the detailed equations of the perturbation term P_k.

Considered together, equations (1,2,3) describe a model of the system. The model has been implemented in Python, allowing us to compute simulations such as the one shown in Figure 2. The first author can provide the code if needed to an interested reader.

In the top plot of the figure, we represent the observable time series belt_k and motor_k from Equation (1) in a double-axis, left for the belt and right for the Z-motor. Observe the different units for each time series. We can see that the (simulated) PI-controller is capable of keeping the system controlled in the presence of the Drift causes, shown in the plot below, as evidenced by the belt position remaining stable around 0. It does this by adjusting the Z-motor position. Eventually, the Z-motor will hit its limit, after which the position of the belt quickly drifts away from its intended position (not shown in the plot). In the bottom plot, we show the different contributors to the drift of the belt, for simplicity of visualization we have combined calibration and misalignment in a single one.

4.2. Inference model

The next step in our methodology is to translate the simulation model into a probabilistic model suitable for inference. In such a model, one describes the unknown variables of the simulation as hidden, i.e. unobserved, random variables. Then one describes the known, i.e. observable, variables as functions of the unknown variables, therefore random variables themselves, but which are observed. These functions relating observable and hidden variables can be: probabilistic (e.g. perturbation), or deterministic (e.g. degradation as a function of s and δ), and need not be invertible. The task of these models is to infer the probability distributions of the hidden random variables that best *explain* the observations. We use the framework of probabilistic programming to instantiate these models and perform inference. Numpyro, see (Phan et al., 2019), is the probabilistic programming language of choice for this work.

Considering the temporal nature of our data and the controlled step-wise nature of the system, we propose a Bayesian state-space model as the probabilistic description. A Bayesian state-space model is a dynamical system of equations relating random variables. The system is determined by the observability and update equations. The observability equation (4) connects the vector of observed variables $\vec{y}_k$ at time k to the vector of hidden variables $\vec{\theta}_k$, external observable variables $\vec{x}_k$ and noise term $\vec{\varepsilon}_k$. The update equation (5) connects the vector of hidden variables at time k with the vector of hidden variables at time $k-1$ and the update noise $\vec{\eta}_k$. Together, they define the system:

$$\vec{y}_k = \mathbf{A}_k\vec{\theta}_k + \mathbf{B}_k \cdot \vec{x}_k + \vec{\varepsilon}_k. \quad (4)$$

$$\vec{\theta}_k = \mathbf{G}_k \cdot \vec{\theta}_{k-1} + \vec{\eta}_k, \quad (5)$$

Where $\mathbf{A_k}$ and $\mathbf{B_k}$ are matrices, $\vec{\varepsilon}_k$ is the observation noise vector at time t, $\mathbf{G_k}$ is often called the innovation or transition matrix at time k and $\vec{\eta}_k$ is the update noise. For this system to be fully Bayesian, we can treat the matrices $\mathbf{A_k}$, $\mathbf{B_k}$,

and $\mathbf{G_k}$, or their coefficients, as random variables themselves and give them Bayesian priors. The equations of a Bayesian state space model describe the evolution of the hidden and observed variables, but not the evolution of their probability distributions. That is the task that the probabilistic program computes in the background.

The translation from a simulation model like that described by equations (1,2,3) into a Bayesian state-space model is not unique and need not be 1-to-1. For instance, the modeler is free to leave elements of the simulation model out of the inference model, implicitly leaving them as contributions to the noise term. They are also free to choose which unknown variables in the simulation model should be mapped to hidden variables in the Bayesian state space model and which should be expressed as coefficients of the matrices $\mathbf{A_k}$, $\mathbf{B_k}$, and $\mathbf{G_k}$, which are treated as static random variables.

For our conveyor belt, we have only one observed variable, $y_k := \mathsf{drift}_k$. The parameters of the model are defined as

$$\mathbf{A} := [1, 0], \quad \mathbf{B} := [c, m], \tag{6}$$

$$\mathbf{G}_k := \begin{bmatrix} 4^\delta & s \cdot 4^\delta - 1 \\ 0 & 1 \end{bmatrix}. \tag{7}$$

The external variables are $\vec{x}_k = [1, \mathsf{print}_k]$, the hidden variables are $\vec{\theta}_k = [D_k, 1]$ and the noise terms are $\varepsilon_k \sim \mathcal{N}(0, \sigma)$ and $\eta_k := 0$.

For simplicity, we choose not to model perturbation P_k explicitly and let it be absorbed by the noise term ε_k. To make this model fully Bayesian, we assign prior distributions to the random variables c, m, δ, s, and σ. We choose uninformative uniform distributions in the interval $[-80, 80]$ for c, m (the whole range of the belt), a wide uniform distributions in $[0, 0.5]$ for δ, uniform 50% prior for each value of s, and a half-normal distribution with width 1 for σ. The ranges of these priors are chosen by domain experts to encompass all plausible failure configurations.

Once we have expressed the model in this form, the probabilistic programming language performs inference by approximating the joint probability distribution of the hidden and observable variables given a particular observation, and applying Bayes rule. Table 1 shows the result of performing inference on the synthetic dataset from Figure 2. Observe how the inferred values for degradation, calibration and misalignment match their real values.

We conclude that the different causes of drift can be inferred from the synthetic data. This gives us reason to believe that, as long as the data from the real system is sufficiently approximated by the simulation model, the inference can also be performed on the real data. This will be shown in the next section.

Table 1. Inferred values for the different sources of drift considered. Errors express a 2-σ confidence interval.

Parameter	Real value	Inferred value
c	15.00	15.02 ± 0.18
m	-20.00	-20.02 ± 0.20
δ	$8.00\,e^{-4}$	$8.00e^{-4} \pm 7e^{-6}$
s	1.00	1.00 ± 0.00

4.3. Results

Given the long service life of the belts, we adapted the inference procedure to better fit the diagnostic needs in the field by inferring the state of the belt at regular intervals. This allows us to track the state of the machine over time. Algorithm 1 outlines the procedure for this periodic computation of degradation, calibration and misalignment given a stream of field data that is split into periods. For each period, the inference engine takes as input the estimated degradation at the beginning of the period, infers the calibration, misalignment and decay exponent, then computes the additional degradation corresponding to that period, and passes the latter to the next iteration.

Algorithm 1 Iterative belt drift inference

```
Data ← [period_1, ..., period_n]
params ← []
prev_D ← 0
for period in Data do
    (c, m, δ, s) ← InferParams(period, prev_D)
    D ← CompDegradation(period, δ, s, prev_D)
    prev_D ← D
    params.append([c, m, δ, s, D])
end for
return params
```

The probabilistic programming comes into play in this algorithm when $params[i] = [c^{(i)}, m^{(i)}, \delta^{(i)}, D^{(i)}]$, the inferred parameters for period i, are treated as probability distributions, rather than point estimates. Inferring these distributions is done through the use of Markov Chain Monte Carlo methods for approximating a total probability distribution. The posterior probability conditioned on the observed data is computed via the Bayes-Laplace rule, rather than a standard parameter fitting technique like minimum square error. This is all handled in the background by the probabilistic programming library and implemented via the function $InferParams$ in the algorithm.

To compute the additional degradation in a given period we modify our hypothesis for the degradation (see eq. (3)) to allow for varying decay exponents over the different periods.

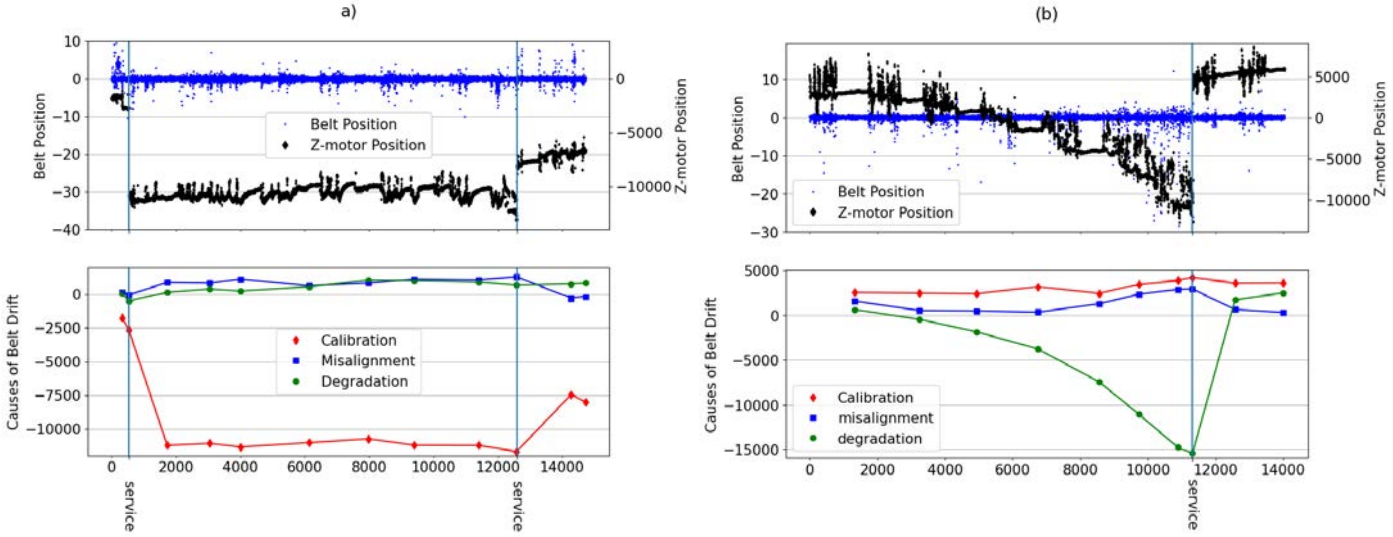


Figure 3. Example of measured data where miscalibration (a), and degradation (b) are the main causes of a belt position error. The Belt and Z-motor positions are measured, while the causes of belt drift in the bottom plots are inferred. The sources of drift are shown here in the units of the Z-position motor rather than the belt for comparison with the former.

We assume that degradation in period i follows the equations:

$$D_k^{(i)} = s \cdot 4^{\delta^{(i)} k} + \mathcal{C}^{(i)}, \quad (8)$$

$$\mathcal{C}^{(i)} = prev_D - s \cdot 4^{\delta^{(i)} k_{i-1}}, \quad (9)$$

where $prev_D = D^{(i-1)} = D_{k_{i-1}}^{(i-1)}$ is the computed degradation at the end of period $i-1$, k_{i-1} is the step that marks the end of period $i-1$, and $\delta^{(i)}$ is the computed decay exponent in period i output by $InferParams$. This condition ensures that degradation grows exponentially and that degradation at the beginning of one period is equal to degradation at the end of the previous period, i.e. it ensures continuity. The computed degradation at the end of period i is then $D^{(i)} := D_{k_i}^{(i)}$.

In Figure 3 we apply the procedure to two typical examples from the field of belts where excessive degradation or miscalibration are the cause of a service action by a service engineer.

Although Figure 3 shows the average degradation, misalignment and miscalibration for each period, we also compute posterior distributions for each parameter, alongside a noise parameter for each period, not shown in the plot. Observe how the three different causes of drift are correctly discriminated and tracked over time. Equipped with these results, a service engineer would perform a replacement of the belt in Figure 3-(b) and a re-calibration of the belt in Figure 3-(a), a much less material and time-consuming action.

5. Conclusions

In this paper, we proposed a methodology for model-based diagnostics of cyber-physical systems leveraging generative and inference models. The generative model is compiled using already available knowledge on failure mechanisms, together with control models, and serves a dual function. On the one hand, it helps validate the expert knowledge on failures, by comparing the results of simulations to data from incidents in the field. On the other hand, it is used to validate the inference models by providing us with a controlled test bench in which to test the ability of the inference model to distinguish the different causes of errors. The inference model is derived from the generative model and is used with field data from real incidents to perform root-cause diagnosis.

We then applied our methodology to the case of an industrial conveyor belt in a closed control loop configuration, with several hidden mechanisms driving the belt out of its desired position. With the proposed methodology we correctly identify the different causes of drift of the belt, thus offering valuable advice for the optimal maintenance action.

To the authors' knowledge, this is the first time probabilistic programming has been used for diagnostics of cyber-physical systems. We believe the present paper proves its utility as a tool for probabilistic modeling and inference in the prognostic and health management domain, opening the door to model-driven and physics-inspired diagnostics. In applying the proposed methodology to the case of an industrial conveyor belt we have identified several aspects for future research. The probabilistic programming framework has a prediction functionality that could be used to make prognostic forecasts of remaining useful life. In future research, we plan on adding this aspect to our methodology. Further, the translation of generative models into inference models is a manual process requiring a certain degree of familiarity with inference and statistical modeling. How to automate, fully or

partially, the translation from simulation to inference models remains an open question. Finally, the proposed methodology has been scoped and tested at a subsystem level, comprising a belt, motor, and control mechanism. How to make such models composable and much larger for system-level diagnosis remains a challenge to be addressed in future research.

ACKNOWLEDGMENT

The research is carried out as part of the Carefree program under the responsibility of TNO-ESI in cooperation with Canon Production Printing. The research activities are supported by the Netherlands Ministry of Economic Affairs and Climate, and TKI-HTSM.

The authors would like to thank Robert Passmann and Peter Kruizinga for providing valuable discussions and comments.

REFERENCES

Bartram, G., & Mahadevan, S. (2015). Probabilistic prognosis with dynamic bayesian networks. *International Journal of Prognostics and Health Management*, *6*(4).

Borth, M., & Barbini, L. (2019). Probabilistic health and mission readiness assessment at system-level. In *Proceedings of the annual conference of the phm society* (Vol. 11).

Cox, M., van de Laar, T., & de Vries, B. (2019). A factor graph approach to automated design of bayesian signal processing algorithms. *International Journal of Approximate Reasoning*, *104*, 185–204.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062.

De Kleer, J., & Kurien, J. (2003). Fundamentals of model-based diagnosis. *IFAC Proceedings Volumes*, *36*(5), 25–36.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*.

Lucas, P. J. (2001). Bayesian model-based diagnosis. *International Journal of Approximate Reasoning*, *27*(2), 99–119.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). *Composable effects for flexible and accelerated probabilistic programming in numpyro.*

Särkkä, S., & Svensson, L. (2023). *Bayesian filtering and smoothing* (Vol. 17). Cambridge university press.

Srinivas, S. (1995). *Modeling techniques and algorithms for probabilistic model-based diagnosis and repair*. stanford university.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., … Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 969–977).

van de Meent, J.-W., Paige, B., Yang, H., & Wood, F. (2018). An introduction to probabilistic programming. *arXiv preprint arXiv:1809.10756*.

van Gerwen, E., Barbini, L., & Nägele, T. (2022). Integrating system failure diagnostics into model-based system engineering. *INSIGHT*, *25*(4), 51–57.

LSTM and Transformers based methods for Remaining Useful Life Prediction considering Censored Data

Jean-Pierre NOOT[1,2]
Etienne BIRMELE[1]
François REY[2]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS*
7 rue René-Descartes, 67000 Strasbourg, France
jnoot@unistra.fr
birmele@unistra.fr
[2] *Liebherr Components Colmar, Haut-Rhin, 68000, FRANCE*
jean-pierre.noot@liebherr.com
francois.rey@liebherr.com

ABSTRACT

Predictive maintenance deals with the timely replacement of industrial components relatively to their failure. It allows to prevent shutdowns as in reactive maintenance and reduces the costs compared to preventive maintenance. As a consequence, Remaining Useful Life (RUL) prediction of industrial components has become a key challenge for condition based monitoring. In many applications, in particular those for which preventive maintenance is the general rule, the prediction problem is made harder by the rarity of failing instances. Indeed, the interruption of data acquisition before the occurrence of the event of interest leads to right censored data. There are few articles in the literature that take that phenomenon into account for RUL prediction, even though it is common in the industrial environment to have a high rate of censored data.

The present article proposes a deep-learning approach based on multi-sensor time series which allows to consider censored data during the training of the neural networks. Two methods are proposed, respectively based on the Dual Aspect Self-Attention based on Transformer proposed by (Z. Zhang, Song, & Li, 2022) for non-censored data and on a recurrent neural network. Their evaluation on the C-MAPSS benchmark dataset shows, compared to the state-of-the-art RUL prediction methods, no loss in the absence of censoring, and outperformance on censored data.

1. INTRODUCTION

Technology and electronic developments of sensors nowadays allow the collection of huge amounts of data on mechanical and industrial equipment, in particular time series measuring their evolution over time. The definition of the maintenance schedule, which is crucial for the industry, therefore shifts to predictive, or condition-based maintenance (CBM). The latter is defined by opposition to the historical preventive maintenance, for which the maintenance schedule is pre-defined, each component being replaced at fixed time intervals. CBM avoids replacement of healthy components, and therefore reduces costs, by determining a dynamic schedule depending on the real-time monitoring of the system. A crucial step is therefore the estimation, given the actual status of the system, of the Remaining Useful Lifetime (RUL) of a component, that is the time before its failure.

Several approaches exist to create CBM models for RUL estimation (Arena, Collotta, Luca, Ruggieri, & Termine, 2021), most of them being model-based methods, data-driven methods or hybridisation of those approaches.

Model-based methods consider the physical phenomenon, for instance corrosion or fatigue, that leads to the failure. A mathematical model is used to simulate the studied mechanism and to get a RUL prediction (Tinga & Loendersloot, 2019). A precise physical and mechanical knowledge is however needed to build physical-based models. Moreover, this approach results in highly complex models when applied to large scale industrial systems composed with a lot of subsystems.

Data-driven methods regroup approaches that rely on stochas-

tic models or statistical analysis to create fault detection models not directly mimicking the underlying physics. It may consist in statistical algorithms to diagnose battery fault (Y. Zhao, Liu, Wang, & Hong, 2017), stochastic processes to mimic the degradation processes (Garay & Diedrich, 2019) or evolving fuzzy models for semiconductor health management (Boutrous, Bessa, Puig, Nejjari, & Palhares, 2022).

Data-driven methods include machine learning algorithms, which have been extensively used by the Prediction and Health Management community to establish predictive maintenance rules. Multiple linear regression durability models were for instance used to predict the fatigue life of automotive coil (Kong, Abdullah, Schramm, Omar, & Haris, 2019), or SVM classifiers for fault detection in vehicle suspensions (Jeong & Choi, 2019). In (Vasavi, Aswarth, Pavan, & Gokhale, 2021), a kNN classifier is used to detect fault by predicting vehicle health using real time data, while (Patil et al., 2018) relies on decision trees and gradient boosting regressor for RUL prediction.

Deep learning, like machine learning methods, allow to have no physical or mechanical knowledge of the studied system. In recent years, numerous articles have demonstrated the effectiveness of those methods for RUL prediction. The data at hand being mainly time series, the developed methods focus on architectures widely used to treat sequential data. Recurrent neural networks like Long-short-time-memory (LSTM) (Zheng, Ristovski, Farahat, & Gupta, 2017), or Convolutional neural network (CNN) (Sateesh Babu, Zhao, & Li, 2016) and recently Transformers (Z. Zhang et al., 2022), which were adapted from the original Transformer architecture (Vaswani et al., 2017) to deal with time series are popular method used to perform RUL predictions.

The presence of right-censored data is an important issue in many real-life industrial applications, which is not taken into account by most methods. Indeed, when the current policy on the field application is predictive maintenance, equipment's are renewed before failure, leading to numerous time-series in the dataset for which the RUL is unknown. One way to deal with such data is to use the survival approach based on Cox models that has been successfully transposed from medical analysis to maintenance analysis (Chen et al., 2020; Yang, Kanniainen, Krogerus, & Emmert-Streib, 2022). An alternative is the ordinal regression (OR) approach where the RUL prediction is replaced by a vector of predictions encoding the failure time (Vishnu, Malhotra, Vig, & Shroff, 2019).

The present paper deals with a new deep-learning method based on ordinal regression to predict RUL on censored data. It relies on two main contributions regarding the state of the art. Firstly, the DAST model (Z. Zhang et al., 2022) based on Transformers is adapted to an ordinal regression framework. Secondly, it is put onto competition with an improved of the LSTM-OR model (Vishnu et al., 2019) to obtain the final prediction rule.

To illustrate its performance, the proposed method is run on the C-MAPSS Turbofan NASA benchmark dataset, and compared to state-of-the-art methods, able to consider censored data or not. The benchmark dataset is being characterized by the absence of censor, the latter is artificially introduced at various levels. The proposed method is comparable to the best methods on non-censored data and better when a significant amount of data is censored.

2. Related Work

As stated in the introduction, the aim of this study is to consider the RUL prediction problem when the learning dataset is right-censored. That situation is common in applications, as such a censoring corresponds to components changed before the failure. This section introduces the main ideas of the DAST (Z. Zhang et al., 2022) and LSTM-OR (Vishnu et al., 2019) architectures, and then builds upon those ideas to propose a novel method for RUL estimation on censored data.

Beforehand, let us introduce the notations which will be used throughout the paper.

For a given unit, we denote by:

- $\mathrm{T}^\star$ the time of failure,
- C the censoring time if relevant, that is if the unit is replaced before failure,
- $\mathrm{T} = \min(\mathrm{C}, \mathrm{T}^\star)$ the observed time of replacement,
- X the time series of the p sensors data, $x_{k,t}$ being the measure of sensor k at time t,
- Z the optional of vector covariates, that is characteristics of the unit which are not varying with time.

Let us fix a maximum value R_{max} for the RUL estimation, which is standard procedure (Heimes, 2008; H. Li, Zhao, Zhang, & Zio, 2020) and is relevant for the applications, as it focuses on the precision of the method on the period preceding the failure. At a given time point t, we then define the lifetime to predict by

$$\mathrm{R}_t = \min(\mathrm{T}^\star - t, \mathrm{R}_{max})$$

Note that this lifetime is observed in the training set only when $\mathrm{T}^\star = \mathrm{T}$. If not, the only available information is that $\mathrm{R}_t \geq \min(\mathrm{C} - t, \mathrm{R}_{max})$.

All the variables in that section are in fact indexed by the number i of the considered unit, for instance when computing a loss. That index is omitted unless necessary for reading purposes.

2.1. Dual Aspect Self-Attention based on Transformer (DA-ST)

The DAST model is an encoder-decoder, with the specificity of a double encoding, using a time step encoder and the sensor encoder.

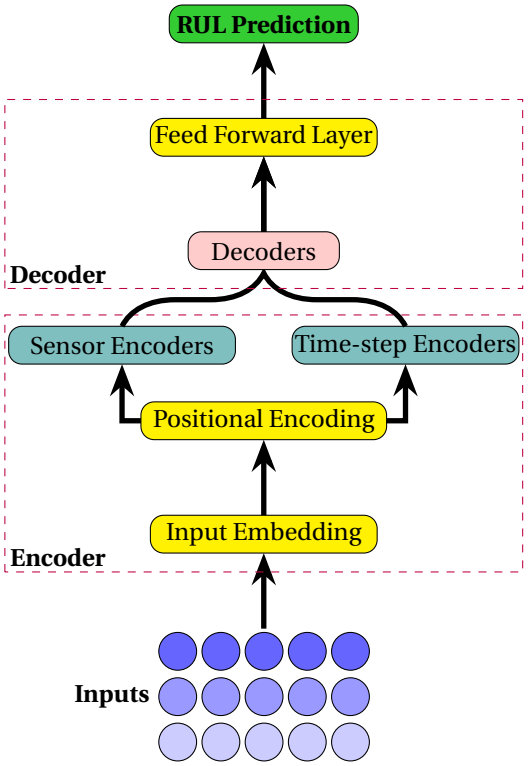


Figure 1. Original DAST architecture (Z. Zhang et al., 2022)

The input data of the DAST architecture consists in a decomposition of the times series X by a sliding window processing of width W, as shown in Figure 2. The input is thus a list of matrices (X_t), each of size (p, W):

$$X_t = \begin{pmatrix} x_{1,t} & \cdots & x_{1,t+W} \\ \vdots & \ldots & \vdots \\ x_{p,t} & \cdots & x_{p,t+W} \end{pmatrix} \quad (1)$$

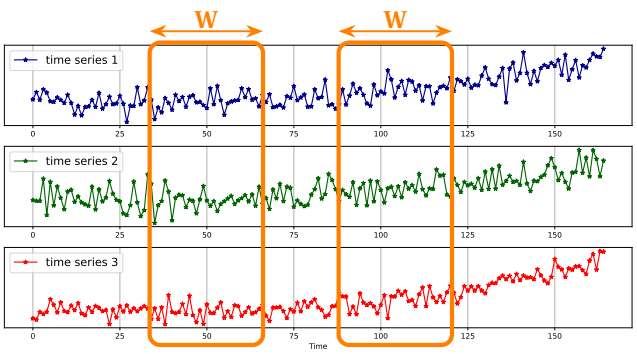


Figure 2. Example of sliding window of size W for time-series on 3 sensors

The data of each window are normalized to equalize the amplitude for each sensor and completed with positional encoding to keep track of the relative time positions of the columns as well as constant lines corresponding to the covariates Z. It is also enriched for each sensor by the mean value and the slope of the linear regression as a function of time, as proposed in (Song et al., 2020).

The originality of DAST is to consider these inputs in two dimensions. On the one hand, the enriched matrix X_t is given as input of the time step encoder, which encodes through self-attention scores per time point the dependency between the vectors of data at different time points. On the other hand, its transpose X_t^T is given as input to the sensor encoder which uses the same architecture to encode and capture the dependency information between the sensors. A final fusion layer finally allows to mix both encodings into a final one, which contains the importance of different combinations of sensors and time steps at the same time. That information is valuable in the context of RUL estimation and is processed by the decoder part of the architecture to obtain a prediction.

As the prediction is a scalar corresponding to $\hat{R}_t$, the model is trained using a RMSE loss, that is the square root of the mean squared prediction error when summing over all units i and time points t.

2.2. Ordinal Regression for RUL Estimation with censored data

In various applications, the complete lifetime of the units is not systematically available as the components may be changed before failure, leading to right-censored lifetime. Direct RUL estimation requires the complete lifetime of the components in the learning data set and thus discards such data, which may represent most of the available data. One possible method to integrate both right-censored and uncensored lifetime data, is the ordinal regression approach developed in (Vishnu et al., 2019).

The key idea is to discretize the object to be estimated, by replacing the RUL $R_t^\star$ by a binary vector of the component status in the future. To do so, two integers L and K are fixed and the status of the unit is checked one time every L cycles (or time points in the time series). The new target is then a vector Y_t of length K where

$$y_{t,k} = \begin{cases} 0 & \text{if} \quad T > t + kL, \\ & \text{i.e. the unit is healthy after } k * L \text{ cycles}, \\ 1 & \text{if} \quad T \leq t + kL \quad \text{and} \quad T = T^\star, \\ & \text{i.e. the unit has failed before } k * L \text{ cycles}, \\ - & \text{if} \quad T \leq t + kL \quad \text{and} \quad T = C, \\ & \text{i.e. the unit status is unknown after } k * L \text{ cycles} \end{cases}$$

t is the time of the current time step and k is the index of Y_t.

Let us for example consider L = 10 and K = 10:

- if the component fails after 75 cycles, $Y_t = (0,0,0,0,0,0,0,0,1,1,1)$,
- if the component is replaced after 75 cycles but before failure $Y_t = (0,0,0,0,0,0,0,0,-,-,-)$. The last three elements are masked as no status appropriate for learning is available.

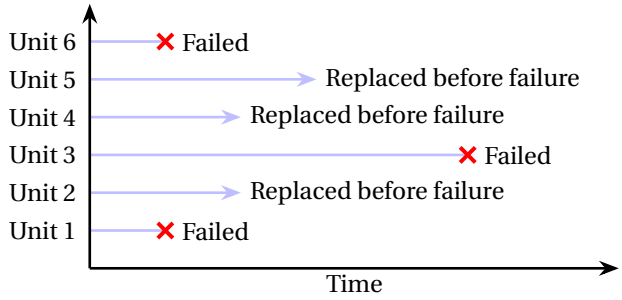


Figure 3. Right-censored data: unit 2, 4 and 5 are censored to the right, they were still healthy when replaced

A learning phase applied on the binary vectors of the training set allows then to obtain a prediction rule, as initially proposed using an LSTM architecture (Vishnu et al., 2019). The prediction for a given unit at time t, denoted by $\hat{Y}_t$, is a vector of K probabilities indicating the probability of failure before the corresponding time steps.

As the problem has become a binary classification problem, the learning is done using the binary cross-entropy (BCELoss). It is however adjusted for right-censored data by discarding all coordinates equal to - in the Y vectors. For example, if $Y_t = (0,0,0,0,0,0,0,0,-,-,-)$, its contribution to the loss is only computed on the seven first coordinates. In other terms, the loss is the sum over all units i and times t of

$$\mathrm{BCE}(Y_t, \hat{Y}_t) = -\sum_{k=1}^{K} \left(y_{t,k} log(\hat{y}_{t,k}) + (1 - y_{t,k}) log(1 - \hat{y}_{t,k}) \right) \tag{2}$$

where the term in the sum is set to 0 whenever $y_{t,k}$ is masked.

2.3. The proposed method

We consider a framework to deal with censored data using the OR encoding with the following step:

1. Adapt the DAST architecture to the OR framework by adding a sigmoid layer, leading to the **DAST-OR** architecture. After training, it outputs a vector $(\hat{Y}_t)$ of length K for every time point in a time series.
2. Following (Chaoub, Voisin, Cerisara, & Iung, 2021) which studies LSTM for RUL prediction, a feed-forward-layer is added in the LSTM-OR architecture, between the LSTM and the sigmoid output layer.This model is denoted as **LSTM-MLP-OR**. It outputs an alternative vector $(\hat{Y}_t)$ of length K for every time point in a time series.
3. Map every vector $\hat{Y}_t$ into a predicted RUL $\hat{R}_t$, following (Vishnu et al., 2019):

$$\hat{R}_t = R_{max}(1 - \frac{1}{K}\sum_{k=1}^{K} \hat{y}_{t,k}) \tag{3}$$

with $R_{max} = KL$ being chosen as the length of the time interval covered by $\hat{Y}_t$.

4. Select the best model in terms of RMSE loss of this RUL prediction on the validation data.

Note that the RUL estimation introduced step 3 is of practical use, but also allows comparison with methods in the literature estimating directly the RUL.

Moreover, to reduce randomness, 10 train of each model are performed, leading to two options:

1. **The simple model:** The model with the best loss on the validation dataset is chosen.
2. **The ensemble model:** We consider an ensemble of models, the final prediction corresponding to the average prediction of the 6 best models among the 10 models trained.

3. Experimental Evaluation

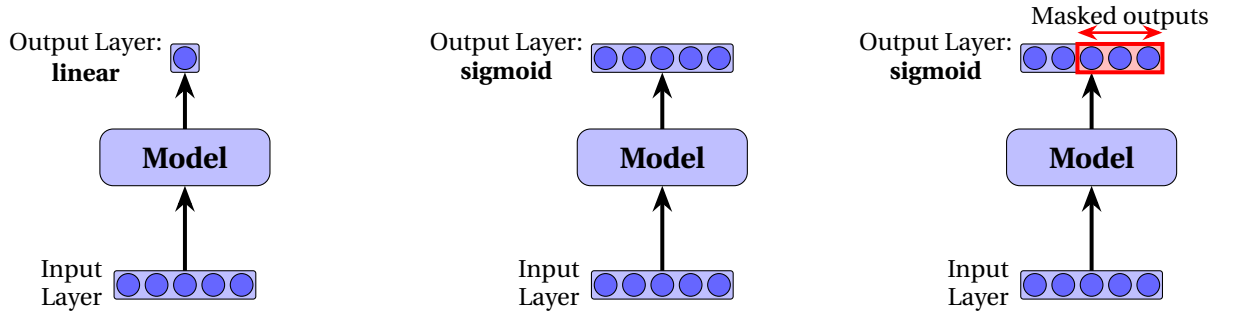
3.1. The CMAPSS dataset

The performance of the proposed method is evaluated on the **C-MAPSS** (Commercial Modular Aero Propulsion System Simulation) dataset, which is used as a benchmark for RUL estimation methods. It simulates run-to-failure trajectories of turbofan engines (Saxena, Goebel, Simon, & Eklund, 2008) in two different operating conditions and two failure modes, leading to four sub-datasets FD001, FD002, FD003 and FD004. The characteristics of the four sets are summarized in Table 1. Each trajectory contains the following variables:

1. a unit number corresponding to the component identifier,
2. a time variable corresponding to the number of cycles performed,
3. the simulation parameters (operating condition and failure modes),
4. the simulated data from 21 sensors.

Table 1. Summary of C-MAPSS dataset

C-MAPSS sub-datasets	FD001	FD002	FD003	FD004
Train trajectories	100	260	100	249
Test trajectories	100	259	100	248
Operating condition	1	6	1	6
Fault modes	1	1	2	2



(a) Metric regression (b) Ordinal regression (c) Ordinal regression with censored data

Figure 4. Metric regression compared to ordinal regression

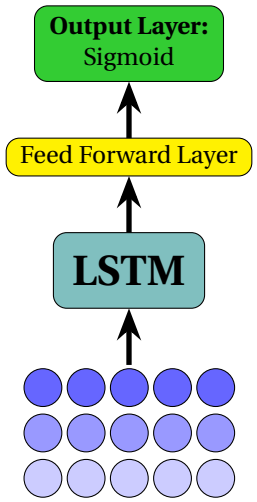


Figure 5. LSTM-MLP-OR architecture

3.2. Data preprocessing and censoring

Sensors having a constant value during the experiment are removed, leaving 14 sensors for datasets FD001 and FD003, and 21 for datasets FD002 and FD004. Data standardization is processed on the remaining sensors by removing the mean and scaling by standard deviation.

Right-censoring is artificially added to the data, with rates $p_c \in [0\%, 20\%, 50\%, 70\%, 90\%]$. More precisely, for every censor rate p_c, the corresponding fraction of the units are randomly chosen, and, for each selected unit, the time series are truncated prior to failure at a random moment. When $p_c = 90\%$, it leads to a train set where only 10% of the units have a known RUL, and approximately 45% of the initial data is kept.

Finally, to be able to chose the best models during the training, each sub-dataset is divided into a training set and a validation set, 20% of randomly chosen units joining the validation set.

3.3. Trained models

Three architectures are trained on the four datasets of the CMAPSS data:

1. **DAST for RUL** (Z. Zhang et al., 2022),
2. **LSTM-MLP-OR**,
3. **DAST-OR**.

Moreover, each of them are trained ten times, and those results are used to derive a single and an ensemble model for each architecture. Ensemble models are denoted with the addition of a final E, for instance DAST-ORE for the ensemble version of DAST-OR.

We also consider the model **BEST-ORE** which is chosen between DAST-ORE and LSTM-MLP-ORE based on the RMSE on the validation dataset.

Seven different models are thus obtained, which can be fairly compared, on exactly the same preprocessing, censoring, as well as training, validation and test sets.

For all models, we use $R_{max} = 130$, and for the methods relying on ordinal regression, we consider Y vectors consisting on K = 13 coordinates corresponding to the status every L = 10 cycles.

3.4. Hyperparameters of the models

The hyperparameters employed for the DAST are those described in the original article (Z. Zhang et al., 2022), except for number of epochs that is set to 250 with early stopping. They are summarized table 2.

The hyperparameters for DAST-OR are essentially identical. Table 3 presents those which are specific to DAST-OR (sigmoid output layer and loss) or are chosen different (a manual tuning on the window size gave better results). The number of epochs is set to 500 with early stopping.

The hyperparameters of the LSTM-MLP-OR model mainly correspond to the article introducing LSTM-OR (Vishnu et al., 2019). However, not all parameters being explicitly de-

Table 2. Hyperparameters of DAST

Hyperparameter	Value
Input embedding	1 MLP layer with 64 neurons, activation: Linear
Sensor encoder	N = 2 Sensor encoder blocks with H = 4 heads
Time step encoder	N = 2 Timestep encoder blocks with H = 4 heads
Decoder	N = 1 Decoder block with H = 4 heads
Output layer	1 MLP layer with 64 neurons, activation: ReLU
Final output layer	1 MLP layer with 1 neurons, activation: Linear
Learning Rate	0.001
Batch Size	256
Dropout	0.2
Window Size	40 for FD001 and FD003, 60 for FD002 and FD004
Optimizer	Rectified Adam
Loss	RMSE

Table 3. Hyperparameters of DAST-OR

Hyperparameter	Value
Final output layer	1 MLP layer with 13 neurons, activation: Sigmoid
Window Size	60 for FD001 and FD003, 80 for FD002 and FD004
Loss	BCELoss

tailed in the original article, manual tuning has been applied to the LSTM-MLP-OR model with a few trials on the validation set.

3.5. Results on uncensored data

This part focuses on the comparison of the results obtained on data without censoring. Table 4 shows the results for the seven trained models on each of the four datasets, with various state-of-the-art methods. Note that the seven methods are trained with the same preprocessing and separation into training and validation sets, whereas the reported values for other methods correspond those indicated in the corresponding publications. Small variations may therefore not be significant.

On FD001, results of DAST-ORE are equivalent to the results of DAST and F+T. On FD002 results of LSTM-MLP-OR and LSTM-MLP-ORE are significantly better than the results obtain with DAST, and equivalent to results obtain with MLP+LSTM and F+T. On FD003 DAST-ORE perform better than other models of the state of the art. The results of LSTM-MLP-ORE are equivalent to the result of MLP+LSTM and F+T. On FD004 LSTM-MLP-ORE perform significantly better than other models. All the OR method proposed are significantly better than the LSTM-ORCE.

Two main conclusions can be drawn from those results. The first is that, even if OR models were designed to handle right-censored data, the obtained results on uncensored data are equivalent to those found in the literature with models specifically made for direct RUL estimation. The second interesting fact to note is the dependence on the number of operating conditions in the dataset (cf Table 1). Differences between sets FD001 and FD003, with a unique operating condition, and sets FD002 and FD004, with six different ones, are commonly found in the state of the art (C. Zhao, Huang, Li, & Yousaf Iqbal, 2020) (Sateesh Babu et al., 2016) (C. Zhang, Lim, Qin, & Tan, 2016) (Zheng et al., 2017) (X. Li, Ding, & Sun, 2018). Furthermore, the number of inputs used between is different. In this study, it appears that DAST-based methods are more powerful when the operating condition is unique, while LSTM-based outperform them when there are 6 operating conditions. Learning both architectures and keeping the best on the validation set, as does BEST-ORE, is therefore useful.

3.6. Results on censored data

The results on the C-MAPSS dataset for each right-censored rate are detailed in Tables 6 and 5. The former compares the proposed ensemble methods to the ensemble LSTM-ORCE method (Vishnu et al., 2019) for the data subsets (FD001 and FD004) and censoring rates studied in that article. The train and validation sets being different, small variations should not be interpreted. However, it clearly indicates a better performance of DAST-ORE on FD001 and a significant improvement with LSTM-MLP-ORE due to the supplementary MLP layer on FD004.

Table 6 shows the results for the models listed in section 3.3 trained on the same training and validation data. For readability, BEST-ORE is not indicated, but the associated RMSE is always the lowest among the RMSEs of LSTM-MLP-ORE and DAST-ORE.

The FD001 dataset has more simple operating conditions and more simple failure modes than the other C-MAPSS sub-datasets. On FD001 the DAST-ORE model has the best RMSE for each percentage of right-censored value. With the increase of p_c, the RMSE is slowly deteriorating and reach it's worst value at $p_c = 90\%$, which is not a surprise as the learning data becomes less informative. Other models, especially LSTM-based ones, show a bigger deterioration with increasing censoring.

The results are similar on FD003, which has also only one operating condition but two failure modes. The best overall results are obtained with DAST-ORE. Moreover, the increasing of the RMSE for highly censored data is milder for DAST-ORE than for competing methods.

FD002 and FD004 are considered more complex than FD001

Table 4. RMSE results on the C-MAPSS dataset without censoring

Model	**FD001**	**FD002**	**FD003**	**FD004**	Average RMSE
LSTM-MLP-OR	14.24	**12.00**	17.27	15.35	14.94
LSTM-MLP-ORE	13.20	12.77	13.84	14.75	13.64
DAST	12.35	16.48	13.43	19.89	15.54
DAST-E	12.22	15.44	12.89	16.14	14.17
DAST-OR	12.16	15.62	9.64	16.20	13.41
DAST-ORE	11.57	15.55	**8.54**	18.01	13.42
BEST-ORE	11.57	12.77	**8.54**	14.75	**11.91**
DAST (Z. Zhang et al., 2022)	**11.43**	15.25	11.32	18.36	14.09
LSTM-ORCE (Vishnu et al., 2019)	14.62	-	-	27.47	-
MLP+LSTM (Chaoub et al., 2021)	13.26	12.49	13.11	**13.97**	13.21
F+T (Lai, Liu, Pan, & Chen, 2022)	**11.43**	13.32	11.47	14.38	12.65

Table 5. Results RMSE on C-MAPSS

	FD001					
p_c	LSTM-MLP-OR	LSTM-MLP-ORE	DAST	DAST-E	DAST-OR	DAST-ORE
0%	14.24	13.20	12.35	12.22	12.16	**11.57**
20%	15.42	14.01	13.69	12.59	12.73	**12.51**
50%	15.09	15.96	15.41	13.37	13.39	**12.99**
70%	17.83	17.97	15.38	14.08	14.28	**12.51**
90%	30.02	26.76	16.78	17.17	17.01	**15.80**
	FD002					
p_c	LSTM-MLP-OR	LSTM-MLP-ORE	DAST	DAST-E	DAST-OR	DAST-ORE
0%	**12.00**	12.77	16.48	15.44	15.62	15.55
20%	15.43	**13.01**	14.09	13.80	16.37	18.51
50%	13.71	**13.15**	15.08	14.18	15.39	16.58
70%	14.24	**13.24**	16.10	14.74	16.71	17.73
90%	16.44	**13.61**	15.85	15.08	25.23	17.00
	FD003					
p_c	LSTM-MLP-OR	LSTM-MLP-ORE	DAST	DAST-E	DAST-OR	DAST-ORE
0%	17.27	13.84	13.43	12.89	9.64	**8.54**
20%	15.69	12.80	13.55	12.53	10.03	**8.81**
50%	13.97	13.46	16.04	12.57	11.69	**10.14**
70%	21.72	21.13	20.88	15.32	13.46	**12.20**
90%	38.74	30.66	22.34	22.88	19.59	**16.09**
	FD004					
p_c	LSTM-MLP-OR	LSTM-MLP-ORE	DAST	DAST-E	DAST-OR	DAST-ORE
0%	16.23	**14.75**	19.89	16.14	16.20	18.01
20%	15.66	**14.42**	18.32	15.23	18.01	16.93
50%	16.00	**14.67**	17.46	15.66	17.43	19.49
70%	16.59	**15.11**	17.32	17.10	14.84	20.83
90%	18.85	**15.47**	19.79	17.21	22.41	22.14

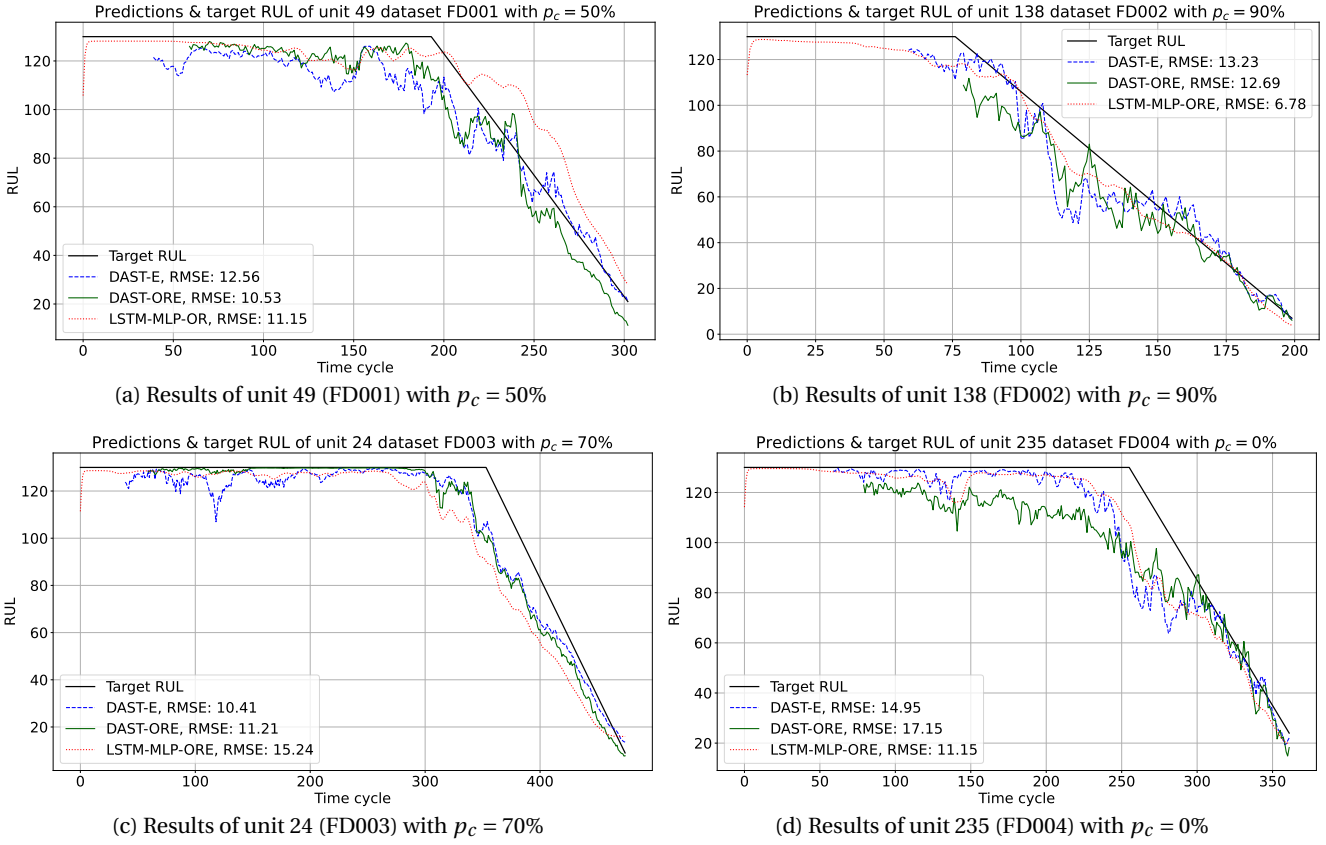
Table 6. RMSE results on FD001-FD004 with censor

p_c	LSTM-MLP-ORE	DAST-ORE	LSTM-ORCE (Vishnu et al., 2019)
	FD001		
50%	15.96	**12.99**	15.98
70%	17.97	**12.51**	16.57
90%	26.76	**15.80**	20.38
	FD004		
50%	**14.67**	19.49	30.62
70%	**15.11**	20.83	31.27
90%	**15.47**	22.14	38.41

and FD003, because they mix several operating conditions. In both cases, the LSTM-based models outperform the DAST-based ones, as for uncensored data, with a small advantage for the LSTM-MLP-ORE ensemble method. In those two cases, the decrease of performance with growing censoring is remarkably low.

The conclusion of this study is therefore two-fold. First, the competition between LSTM and DAST-based architectures remains relevant with censored data, as different conditions may lead to different rankings of those methods. Second, OR-based methods allow to obtain a reasonable loss of performance when the real time of failure is missing for most of the learning data.

As prescribed in (Saxena et al., 2008), the results were evalu-



(a) Results of unit 49 (FD001) with $p_c = 50\%$

(b) Results of unit 138 (FD002) with $p_c = 90\%$

(c) Results of unit 24 (FD003) with $p_c = 70\%$

(d) Results of unit 235 (FD004) with $p_c = 0\%$

Figure 6. Example of results on one units of each sub-dataset

ated by the RMSE on the predictions of the last time-point of the time-series in the test set. To illustrate more visually the results of the different methods, Figure 6 provides some plots of the predictions of the ensemble methods for randomly picked time series on different datasets and censoring rates.

3.7. Asymmetric score evaluation

Prediction on C-MAPSS should also be evaluated by the asymetric score (Saxena et al., 2008) defined by

$$\text{Score} = \begin{cases} e^{\frac{\hat{\text{R}}_t - \text{R}_t}{13}} - 1 & \text{if } \hat{\text{R}}_t - \text{R}_t \geq 0 \\ e^{-\frac{\hat{\text{R}}_t - \text{R}_t}{13}} - 1 & \text{if } \hat{\text{R}}_t - \text{R}_t < 0 \end{cases}$$

That score corresponds to a higher penalty for overestimation rather underestimation of the RUL.

Table 7 shows the scores for the ensemble methods DAST-E, DAST-ORE and LSTM-ORE. If the results are rather coherent with the RMSE comparisons for FD001 and FD003, the advantage of LSTM-based methods compared to DAST-E is less clear for FD002 and FD004.

However, it has to be noted that the OR-based methods were trained with a symmetric BCELoss which does not take into consideration a different penalty for over and under-estimations. In terms of binary vectors, it means a higher penalty for a close to 0 coordinate in $\hat{\text{Y}}_t$ when the truth is 1 (the fan is predicted running when it actually failed, which is an over-estimation of the RUL) than for a prediction close to 1 when the truth is 0.

A possibility to introduce this asymmetry would be to consider a modified loss by replacing Equation 2 by

$$\text{BCE}(\text{Y}_t, \hat{\text{Y}}_t) = -\sum_{k=1}^{\text{K}} \left(\alpha y_{t,k} log(\hat{y}_{t,k}) + (1 - y_{t,k}) log(1 - \hat{y}_{t,k})\right) \tag{4}$$

where $\alpha > 1$ is a hyperparameter to be optimized.

4. Conclusion

This work addresses the challenge of estimating the Remaining Useful Life (RUL) of industrial components from time series data with no prior physical model of the system and a high rate of censored data. It does so by considering two data-driven deep-learning architectures relying on the ordi-

Table 7. Score results on C-MAPSS

	FD001		
p_c	LSTM-MLP-ORE	DAST-E	DAST-ORE
0%	341.00	**201.17**	206.48
20%	410,63	**232,49**	269,68
50%	631,91	**252,89**	296,04
70%	1279,51	458,45	**261,79**
90%	4566,18	808,5	**500,21**
	FD002		
p_c	LSTM-MLP-ORE	DAST-E	DAST-ORE
0%	708,28	**638,34**	978,48
20%	753,57	**531,06**	1891,21
50%	751,84	**544,1**	1412,07
70%	786,58	**647,39**	1362,25
90%	860,19	**849,25**	1286,1
	FD003		
p_c	LSTM-MLP-ORE	DAST-E	DAST-ORE
0%	322,06	206,94	**103,65**
20%	250,9	192,28	**111,17**
50%	267,75	223,95	**143,29**
70%	1611,57	437,85	**272,35**
90%	3100,45	2797,84	**447,64**
	FD004		
p_c	LSTM-MLP-ORE	DAST-E	DAST-ORE
0%	**1741,67**	2262,98	2739,62
20%	1772,3	**1518,44**	2591,59
50%	**1434,33**	2206,4	2788,02
70%	**2096,04**	2470,12	3863,9
90%	1689,67	**1194,16**	2903,72

nal regression approach introduced in (Vishnu et al., 2019) for RUL estimation. One of them is an improved version of the LSTM-OR method by (Vishnu et al., 2019), the second is an adaptation to censored data of the DAST model introduced in (Z. Zhang et al., 2022).

These approaches are shown to perform as well on the C-MAPSS data as the existing direct RUL estimation methods found in the literature on uncensored data, and better on censored data.

Furthermore, the two proposed architectures are shown to be complementary, as they outperform each other depending on the complexity of the dataset. Therefore, in the context of estimating the lifespan of components, it is interesting to put them in competition, considering that this approach should yield favorable results regardless of the complexity of the data and the rate of right-censored data.

CODE AVAILABILITY

The code was written in Pytorch and is available at https://gitlab.math.unistra.fr/jnoot/rul_estimation_cmapss

REFERENCES

Arena, F., Collotta, M., Luca, L., Ruggieri, M., & Termine, F. G. (2021). Predictive maintenance in the automotive sector: A literature review. *Mathematical and Computational Applications*, *27*(1), 2.

Boutrous, K., Bessa, I., Puig, V., Nejjari, F., & Palhares, R. M. (2022). Data-driven prognostics based on evolving fuzzy degradation models for power semiconductor devices. In *Phm society european conference* (Vol. 7, pp. 68–77).

Chaoub, A., Voisin, A., Cerisara, C., & Iung, B. (2021). Learning representations with end-to-end models for improved remaining useful life prognostics. *arXiv preprint arXiv:2104.05049*.

Chen, C., Liu, Y., Wang, S., Sun, X., Di Cairano-Gilfedder, C., Titmus, S., & Syntetos, A. A. (2020). Predictive maintenance using cox proportional hazard deep learning. *Advanced Engineering Informatics*, *44*, 101054.

Garay, J. M., & Diedrich, C. (2019). Analysis of the applicability of fault detection and failure prediction based on unsupervised learning and monte carlo simulations for real devices in the industrial automobile production. In *2019 ieee 17th international conference on industrial informatics (indin)* (Vol. 1, pp. 1279–1284).

Heimes, F. O. (2008). Recurrent neural networks for remaining useful life estimation. In *2008 international conference on prognostics and health management* (pp. 1–6).

Jeong, K., & Choi, S. (2019). Model-based sensor fault diagnosis of vehicle suspensions with a support vector machine. *International Journal of Automotive Technology*, *20*, 961–970.

Kong, Y., Abdullah, S., Schramm, D., Omar, M., & Haris, S. (2019). Development of multiple linear regression-based models for fatigue life evaluation of automotive coil springs. *Mechanical Systems and Signal Processing*, *118*, 675–695.

Lai, Z., Liu, M., Pan, Y., & Chen, D. (2022). Multi-dimensional self attention based approach for remaining useful life estimation. *arXiv preprint arXiv:2212.05772*.

Li, H., Zhao, W., Zhang, Y., & Zio, E. (2020). Remaining useful life prediction using multi-scale deep convolutional neural network. *Applied Soft Computing*, *89*, 106113.

Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, *172*, 1–11.

Patil, S., Patil, A., Handikherkar, V., Desai, S., Phalle, V. M., & Kazi, F. S. (2018). Remaining useful life (rul) prediction of rolling element bearing using random forest and gradient boosting technique. In *Asme international mechanical engineering congress and exposition* (Vol. 52187, p. V013T05A019).

Sateesh Babu, G., Zhao, P., & Li, X.-L. (2016). Deep convolutional neural network based regression approach for estimation of remaining useful life. In *Database systems for advanced applications: 21st international conference, dasfaa 2016, dallas, tx, usa, april 16-19, 2016, proceedings, part i 21* (pp. 214–228).

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).

Song, Y., Gao, S., Li, Y., Jia, L., Li, Q., & Pang, F. (2020). Distributed attention-based temporal convolutional network for remaining useful life prediction. *IEEE Internet of Things Journal*, *8*(12), 9594–9602.

Tinga, T., & Loendersloot, R. (2019). Physical model-based prognostics and health monitoring to enable predictive maintenance. *Predictive Maintenance in Dynamic Systems: Advanced Methods, Decision Support Tools and Real-World Applications*, 313–353.

Vasavi, S., Aswarth, K., Pavan, T. S. D., & Gokhale, A. A. (2021). Predictive analytics as a service for vehicle health monitoring using edge computing and ak-nn algorithm. *Materials Today: Proceedings*, *46*, 8645–8654.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vishnu, T., Malhotra, P., Vig, L., & Shroff, G. (2019). Data-driven prognostics with predictive uncertainty estimation using ensemble of deep ordinal regression models. *International Journal of Prognostics and Health Management*, *10*(4).

Yang, Z., Kanniainen, J., Krogerus, T., & Emmert-Streib, F. (2022). Prognostic modeling of predictive maintenance with survival analysis for mobile work equipment. *Scientific Reports*, *12*(1), 8529. Retrieved from https://doi.org/10.1038/s41598-022-12572-z doi: 10.1038/s41598-022-12572-z

Zhang, C., Lim, P., Qin, A. K., & Tan, K. C. (2016). Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE transactions on neural networks and learning systems*, *28*(10), 2306–2318.

Zhang, Z., Song, W., & Li, Q. (2022). Dual aspect self-attention based on transformer for remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–11.

Zhao, C., Huang, X., Li, Y., & Yousaf Iqbal, M. (2020). A double-channel hybrid deep neural network based on cnn and bilstm for remaining useful life prediction. *Sensors*, *20*(24), 7109.

Zhao, Y., Liu, P., Wang, Z., & Hong, J. (2017). Electric vehicle battery fault diagnosis based on statistical method. *Energy Procedia*, *105*, 2366–2371.

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In *2017 ieee international conference on prognostics and health management (icphm)* (pp. 88–95).

Maintenance decision-making model for gas turbine engine components

Hongseok Kim[1], and Do-Nyun Kim[2]

[1]*Department of Mechanical Engineering, Seoul National University, Seoul 08826, Republic of Korea*
saint4561@snu.ac.kr

[2]*Department of Mechanical Engineering, Institute of Advanced Machines and Design, and Institute of Engineering Research, Seoul National University, Seoul 08826, Republic of Korea*
dnkim@snu.ac.kr

ABSTRACT

When designing gas turbine engine components, the inspection and maintenance (I&M) plan is prepared using the safe life. However, the I&M plan determined using safe life may be costly since all components are replaced at designated life. Therefore, it is important to make maintenance decisions considering the time-dependent deterioration process of gas turbine engine components for a cost-saving I&M plan. In this study, we proposed a maintenance decision-making model for gas turbine engine components based on a partially observed Markov decision process (POMDP). Using dynamic Bayesian networks, a decision-making model integrating a reliability analysis model, and a decision model for I&M planning was constructed. The signal amplitude data resulting from non-destructive inspection according to operation hour was used as partially observed data. The total cost obtained from the proposed model were compared with the results using a fixed I&M plan. The proposed model resulted in more cost-effectiveness I&M planning within affordable risk levels by considering the interaction between risk cost and I&M cost.

1. INTRODUCTION

Ensuring the safety of the gas turbine engine is very important in aircraft operation. There are two traditional inspection & maintenance (I&M) strategies to operate aircraft safely; safe life and damage tolerance design. The safe life method (C. H. Cook et al., 1982) replaces all components after the design allowable life, and time-based maintenance (TBM) (Bousdekis et al., 2015) inspects and repairs all parts at predetermined intervals. However, traditional I&M methods require high costs since I&M actions are planned without the consideration of the components' condition. For this reason, a condition-based maintenance method that emphasizes combining data-driven reliability models with condition-monitored data was developed (Alaswad & Xiang, 2017).

Markov decision process (MDP) is one of the widely used methodologies for decision-making models with the condition-based maintenance (CBM) method. MDP takes actions at each stage to maximize the reward under perfect observation of components state. However, there are limitations for MDP that perfect observation of the components state is unrealistic (Papakonstantinou & Shinozuka, 2014b). Partially observable MDP (POMDP) quantified the uncertainty of imperfect observation by estimating the belief of state from the information obtained with the probability of observation. (Papakonstantinou & Shinozuka, 2014b, 2014a) determined the optimal life-cycle policy of concrete structures by implementing the POMDP. (Memarzadeh et al., 2014) proposed the algorithm for approximate learning and planning the Bayes-adaptive POMDP (BA-POMDP) framework to find the optimal maintenance plan of wind farms.

Morato et al. (Morato et al., 2020) incorporated the dynamic Bayesian networks (DBNs) and POMDP to obtain optimal I&M strategy for deteriorating structure. They modeled the deterioration model based on time-invariant parametric DBNs, and an optimal I&M plan was generated by minimizing the total cost of inspection, maintenance, and reliability. Hlaing et al. (Hlaing et al., 2022) presented the non-stationary policy for offshore wind tubular joints by integrating the Bayesian networks and POMDP. They estimated the probability of failure (POF) using the DBNs and obtained optimal I&M policy via POMDP.

In this work, we proposed the maintenance decision-making model for gas turbine engine components. DBNs and POMDP were integrated to get optimal I&M policy. The fatigue crack growth model was implemented for the

deterioration of gas turbine engine components, probability of detection (POD) curve which is a function of the crack size was used for the inspection model. The remainder of this paper is organized as follows. The decision-making model for gas turbine engine components is described in "Methodology" section. In "Numerical results", the optimal I&M policy obtained using the proposed POMDP model is presented. In the final section, the conclusions of this study are summarized.

2. METHODOLOGY

2.1. PARTIALLY OBSERVABLE MARKOV DECISION PROCESS

MDP provides the framework that finds the optimal policy for sequential decision-making problems, as represented in Fig. 1. In Fig. 1, the circles are random state nodes, the rectangular are decision nodes, and the polygons are reward nodes. MDP determines the optimal policy that maximizes the expected reward value by using the Bellman equation as follows:

$$V^*(s) = \max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} T(s,a,s')V^*(s')] \quad (1)$$

where s is states of the system, a is the set of possible actions, and $T(s, a, s')$ is transition matrix which is the probability of transit from current state s_t to next state s_{t+1}, $R(s, a)$ is the reward when doing action a with current state s_t, and γ is discount factor employed when the problem is infinite horizon planning case (Morato et al., 2019). However, since MDP has the limitation of perfect observation, POMDP determines the optimal policy according to the belief state estimated from imperfect observation. In Fig. 2, the belief state s_t is updated from the information of component state obtained at the inspection node z_t, The optimal policy is determined at POMDP as:

$$V^*(s) = \max_{a \in A}\left[b(s)R(s,a) + \gamma \sum_{z \in Z} P(z|b,a)V^*(b')\right] \quad (2)$$

where z is observation, and b is belief state of the component. The belief state b with action m at stage n is ${}^m b^n = b^n \times A_m$, the belief state b^{n+1} at the stage n+1 is updated with degradation model D; $b^{n+1} = {}^m b^n \times D$ (Faddoul et al., 2013).

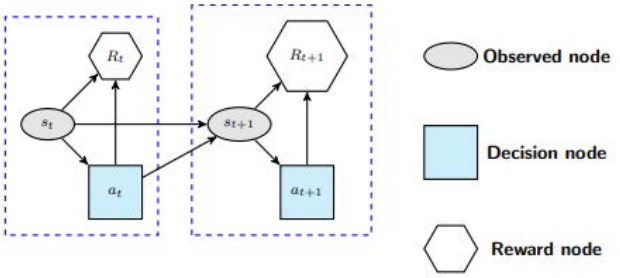


Figure 1. Graphical model for Markov decision process

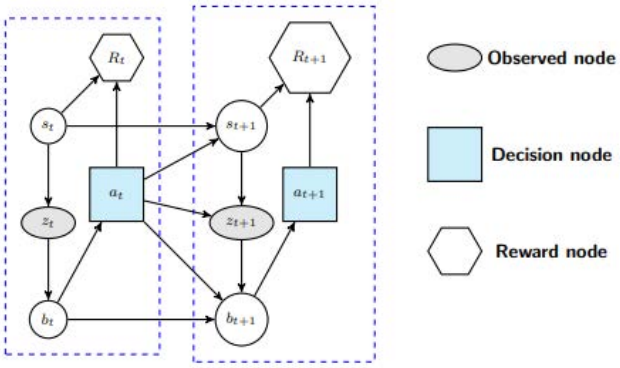


Figure 2. Graphical model for Partially observable MDP

2.2. DETERIORATION AND INSPECTION MODEL

Paris' law was used for fatigue crack deterioration model of gas turbine engine component as:

$$\frac{da}{dN} = C(\Delta K)^m \quad (3)$$

where a is the size of crack, N is the cycles of loads, ΔK is the stress intensity factor range which is function of crack size, shape, and stress range $\Delta\sigma$, and C and m are the constants related to material property.

Eddy current inspection (ECI), one of the non-destructive inspection (NDI) methods, was used as partial observation model to update the belief state of the gas turbine engine component. The POD of ECI depends on the crack size and the detection threshold (Hlaing et al., 2022). The size of the crack is estimated from the ECI signal amplitude in Eq. (4), and the POD is calculated from Eq. (5). Figure 3 presents the relation curve between the signal amplitude and the crack length, and Fig. 4 is the probability of detection (POD) estimated from the detection result data of NDI personnel (D. Lee & Kwon, 2023). The $a_{50/95}$ = 1.123 in Fig. 4 means that the detectable crack size at a 50% probability with 95% confidence is 1.123mm. The size of the crack is estimated

from the ECI signal amplitude in Eq. (4), and the POD is calculated from Eq. (5).

$$\hat{a} = \beta_0 + \beta_1 a + \varepsilon \quad (4)$$

$$POD(a) = \frac{a a^{\gamma}}{1 + a a^{\gamma}} \quad (5)$$

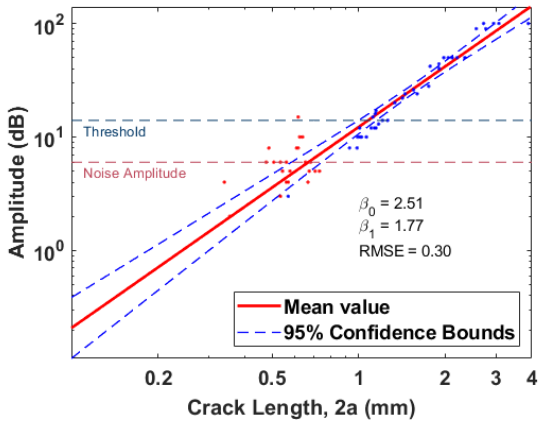


Figure 3. Relation between ECI signal and crack size

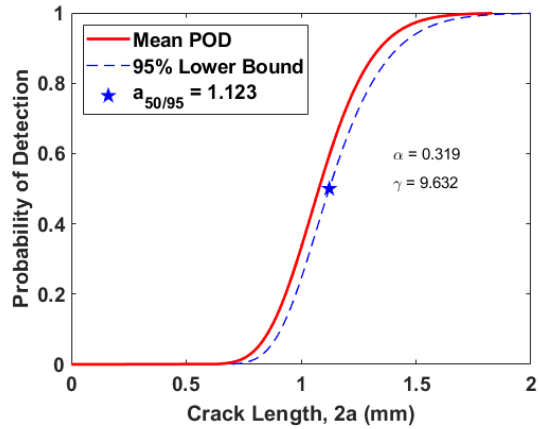


Figure 4. Probability of detection curve

2.3. DECISION-MAKING MODEL BASED ON DYNAMIC BAYESIAN NETWORKS

The maintenance decision-making model based on Dynamic Bayesian Networks (DBNs), depicted in Fig. 5 was developed by incorporating the deterioration, inspection model, and POMDP described above. In Fig. 5, the initial nodes have no parent nodes, the static nodes are time-independently invariant nodes, the observed node obtains evidence, the functional nodes formulate the crack length distribution and reliability, the decision nodes decide for actions, and the cost of the decision incurred in the utility nodes. The continuous operation time is discretized into time slices with uniform intervals.

Table 1. Prior probability distributions of initial nodes

Var	Distribution	Mean	SD	Corr.
m, $\ln(C)$	Binormal	(2.5, $\log(5.2\times10^{-12})$)	(0.3, 0.47)	-0.9
$\Delta\sigma$ (MPa)	Normal	40	5	-
a_0 (mm)	Lognormal	-1.0	0.001	-
a_r (mm)	Lognormal	-1.0	0.001	-
Y_n (%)	Normal	3.29	2.86	-

First, the crack length distribution at time slice t-1 (a_{t-1}) is estimated in the deterioration model using $\Delta\sigma$, m, initial crack length at time slice t-1 (a^0_{t-1}). Next, a_{t-1} is updated to a^*_{t-1} based on the actions determined by the signal amplitude node Y_{t-1}, noise amplitude Y_n, decision nodes for inspection DZ, the threshold of inspection D_{th}, and maintenance DM. The updated crack length distribution at time slice t-1 a^*_{t-1} is used as initial crack length distribution a^0_t at time slice t. The prior distributions of initial nodes are presented in Table 1.

The actions determined in each decision node are as follows: The inspection decision determines whether to perform an inspection or not. The cost of the inspection is obtained in inspection utility node UZ depending on the result of the inspection decision.

- No-inspection: the crack length states transit according to the deterioration model.
- Inspection: binary inspection result is obtained at node Z as 'detected' or 'not detected'. When the inspection result is 'detected', the probability of failure increases as the belief state of crack length larger than the inspection threshold increase. On the other hand, the probability of failure decreases since the crack length distribution smaller than the inspection threshold rise in the case of 'not detected'.

The quality of NDI is determined in the threshold decision node. The inspection quality is high as the threshold is lower. If the signal amplitude obtained at the cracks is larger than the inspection threshold, those cracks are detected at inspection result node Z. There is no cost for threshold decision.

There are binary options in the maintenance decision node; repair or do-nothing. The maintenance utility node UM calculates the cost of maintenance.

- Do-nothing: there is no maintenance action planned in this case, the crack length state evolves according to the stochastic deterioration process.
- Repair: perfect maintenance action is performed. The crack length distribution a^*_{t-1} is replaced by the belief state of repair crack a_r.

The total cost at time t is calculated by summing the cost of the failure, inspection, and maintenance determined according to the results of each decision node as follows:

$$C_T(h) = \sum_{t=t_0}^{t_n} \left[C_I(t)\gamma + C_M(t)\gamma + P_f(t)C_f(t)\gamma \right] \qquad (6)$$

where C_T is total cost, h is pre-defined heuristic schedule, t_n is total time horizon, C_I is inspection cost, C_M is maintenance cost, P_f is probability of failure estimated at node R, and C_f is failure cost determined in utility node UR. The C_I and C_M is not incurred in the case of no-inspection and do-nothing, respectively. The optimal actions were determined by minimizing the total cost.

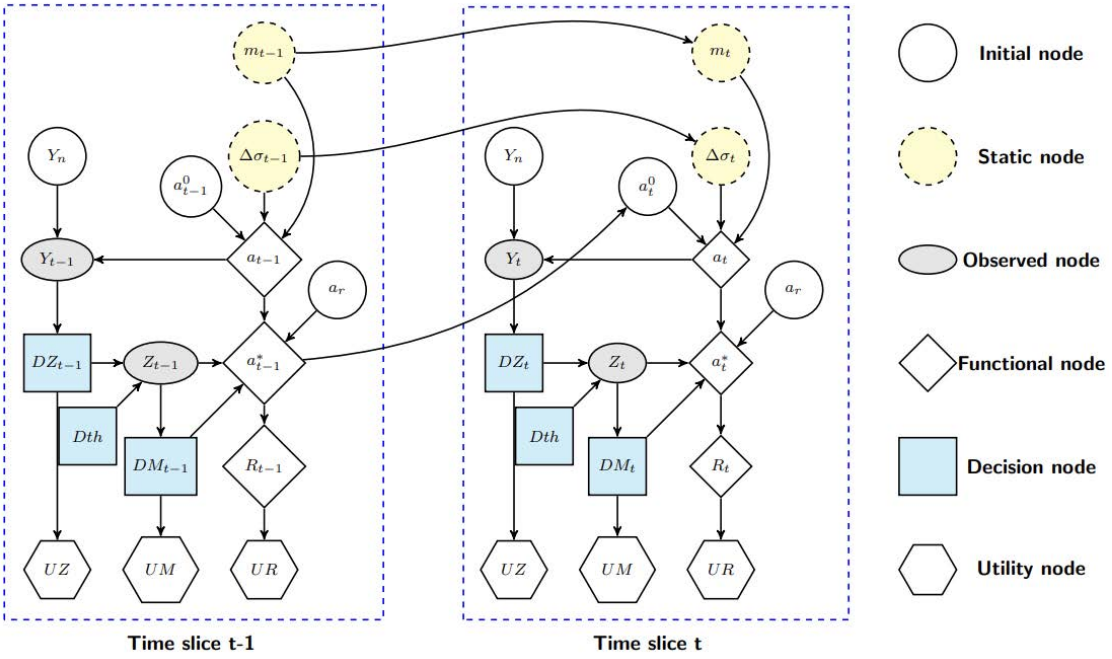


Figure 5. Maintenance decision-making model for gas turbine components

3. NUMERICAL RESULT

A maintenance decision-making model based on POMDP was constructed for the J85 gas turbine engine compressor first-stage rotor blade for the F-5 aircraft. The J85 gas turbine engine compressor first-stage rotor blade is mounted with a disc using tangs. The stress concentration at the center tang occurred due to contact force between the retainer pin and an inner surface of the tang (B. W. Lee et al., 2011). Since the fracture at the center tang may occur due to the fatigue crack initiated from fretting damages by contact stress, it is important to optimize the I&M planning of the blade center tang.

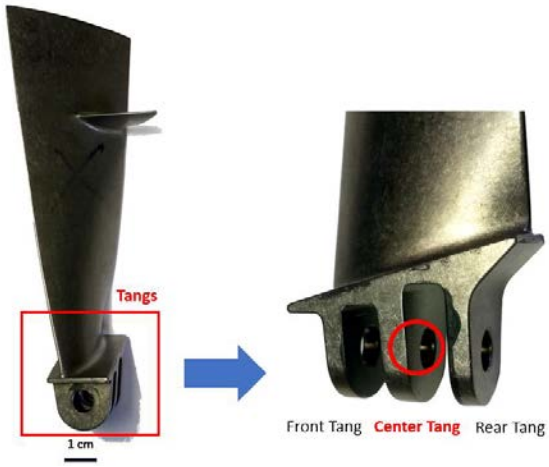


Figure 6. J85 engine compressor first-stage blade

3.1. DISCRETIZATION SCHEMES FOR MODEL PARAMETERS

Discretization of each random variable is necessary in POMDP since the probability of partial observation is discretized (Morato et al., 2020). The accuracy of POF and computational efficiency are affected by the discretization scheme. The discretization schemes for each variable in Fig. 5 are presented in Table 2. The random variables were discretized with a small number of discretization states for computational costs.

Table 2. Discretization schemes

Var	Interval Boundaries
a, a_r	0, exp(log(0.01):(log(3)-log(0.01))/4:log(3)), inf
m	0, log(exp(1):(exp(3.9)-exp(1))/4:exp(3.9)), inf
$\Delta\sigma$	1:60/4:60, inf
Y_n, Y	0:6:100

3.2. OPTIMAL POLICY BASED ON DECISION-MAKING MODEL

The overall cost of utilizing the proposed decision-making model was compared with that of time-based maintenance (TBM), a traditional I&M strategy. In TBM, NDI is conducted for every time slice, and when a crack is detected, a perfect repair action is performed. On the other hand, the decision to execute NDI and repair is made for each time slice in the most cost-effective way in the decision-making model. The parametric study for the costs of inspection, maintenance, and failure was conducted to specify the effects of actions. The state of the measured signal amplitude Y was imported from the actual measured data at each time slice (D. Lee & Achenbach, 2016). If the inspection threshold is smaller than the measured signal amplitude, it is observed that a crack is detected, and a perfect repair action is performed in the TBM strategy. Otherwise, in the POMDP strategy, it is determined whether to perform inspection and maintenance actions for each time slice depending on the total cost.

The ratio of total cost between TBM and POMDP depending on the NDI threshold over 5 time slices is shown in Fig. 7. The evidence indicated the crack state progressed from state 2 to state 5 in each time slice, with failure occurring at state 6. After repair action, the crack state returned to state 2. R_{MI} is the ratio of cost between inspection and maintenance, R_{FM} is the ratio of cost between failure and maintenance, and R_C is the total cost ratio between TBM and POMD as:

$$R_{MI} = \frac{C_M}{C_I}, R_{FM} = \frac{C_F}{C_M} \tag{7}$$

$$R_C = 100\frac{C_{POMDP} - C_{TBM}}{C_{TBM}}(\%) \tag{8}$$

Where C_M is the cost of maintenance, C_I is the cost of inspection, C_F is the cost of failure, C_{POMDP} is total cost for POMDP, and C_{TBM} is that of TBM. The R_{MI} = [10, 20, 30, 40, 50], and R_{FM} = [100, 50, 25, 20] were used to estimate the total cost. The C_{POMDP} is more cost-effective than C_{TBM} when the total cost ratio is a negative value; conversely, if this ratio is a positive value, the C_{TBM} is less expensive than C_{POMDP}.

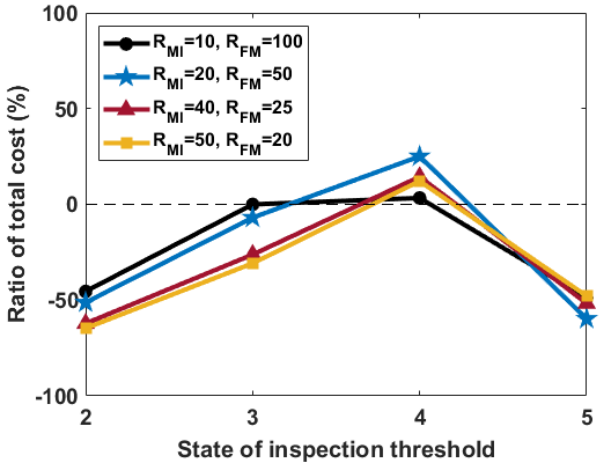


Figure 7. Ratio of total cost between TBM and POMDP

The cost of the POMDP strategy is cheaper than TBM in all R_{MI} and R_{FM} when the threshold of inspection is lower than 4. This implies that a high-quality inspection is crucial for an I&M strategy based on POMDP. When the inspection quality is high (D_{th} = 2, 3), the crack state is identified early. This enables decisions on whether to perform inspections and repairs based on the state of crack growth. Consequently, the POMDP strategy conducts fewer inspections and repairs compared to the TBM strategy, and no preventative repairs, resulting in lower total costs as illustrated in Fig. 8.

When D_{th} = 4, the crack state is detected before it grows near the limit state. In the TBM strategy, after detecting the crack, repairs are performed to maintain a low probability of failure. The inspections following repair prevent the crack from propagating toward the limit state, resulting in a low cost of failure. On the other hand, in the POMDP strategy, when the R_{MI} is low, more repairs are carried out than TBM due to lower inspection quality. When R_{MI} is high, the probability of failure increases because inspections are not performed after repairs, leading to a higher total cost.

Also, the ratio of total cost R_c increases as the inspection threshold decreases. Exceptionally, when using the TBM strategy with D_{th}=5, R_c increases. This occurs as the crack is detected in state 5, which is proximate to the failure state, as depicted in Figure 8(a). Consequently, the estimated probability of failure is relatively high, leading to an increased total cost. On the other hand, when using the POMDP strategy, repairs were carried out preventatively

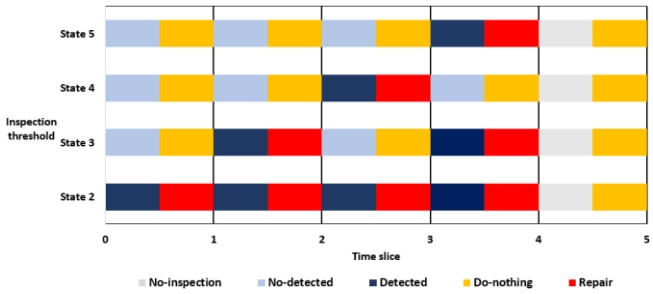
even if the inspection result indicated 'No-detected' as shown in Fig. 8(b), owing to the limited inspection quality. Therefore, when the inspection threshold is 5, the total cost of the POMDP strategy was relatively lower than that of TBM since the failure cost of the POMDP strategy was low through preventive repairs.

Fig. 8 presents the optimal actions determined by using TBM and POMDP strategy depending on each R_{MI} and R_{FM} during 5 time slices. There are two color blocks to describe the results of the decision at each time slice; the left is inspection, and the right is maintenance. The action types of inspection are 'No-inspection' (gray colors), 'No-detected' (sky colors), and 'Detected' (blue colors). The red colors mean the case of a 'Repair' action, and the orange colors indicate a 'Do-nothing' action.

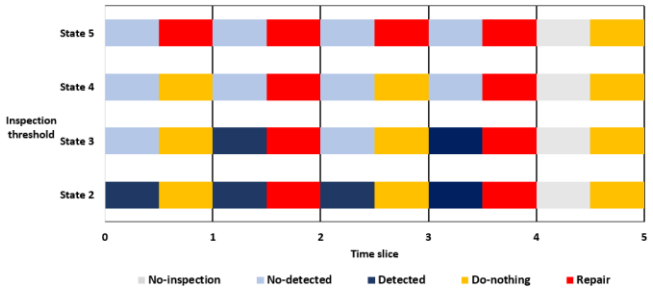
In the context of inspection and maintenance, Fig. 8 illustrates how the cost ratio impact the frequency of inspection and maintenance. Specifically, when the cost ratio of inspection and maintenance R_{MI} is relatively small compared to the cost of failure and maintenance R_{FM} (Fig. 8(b), (c)), more frequent inspection were performed. Since the cost of the failure is expensive compare to inspection and maintenance, it is more cost-effective to identify the state of the crack length early by inspecting frequently. For example, the optimal decision for inspection of R_{MI}=10, and R_F=100 was to inspect every time slice, similar to the TBM strategy.

In the case of high R_{FM}, and high D_{th}, repair action was performed even in the case of 'No-detected'. Since the cost of maintenance is cheaper than that of failure, and the result of inspection is uncertain, this policy is optimal to reduce the POF. On the other hand, when the information quality of inspection was high ($D_{\text{th}}\leq 3$), repair action was not performed immediately, even though the result of the inspection was 'Detected'. In this case, the decision to repair or not can be determined by the condition of the crack, not preventatively.

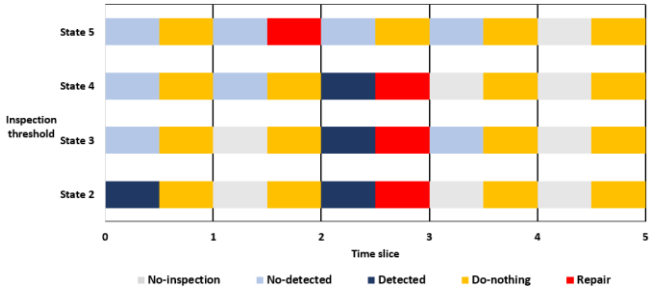
When R_{MI} is larger than R_{FM} (Fig. 8(d), (e)), the cost of maintenance becomes expensive. The preventive inspections and maintenance were reduced due to high-cost maintenance. Therefore, the inspection was not performed at the first time slice for all inspection thresholds. The total cost ratio R_c was highest when R_{MI}=50, R_{FM}=20, and D_{th} =2, as presented in Fig. 7. The findings from Figures 7 and 8 indicate that an increase in inspection quality and a decrease in the cost ratio between maintenance and repair enhance the effectiveness of the maintenance decision-making model based on POMDP.



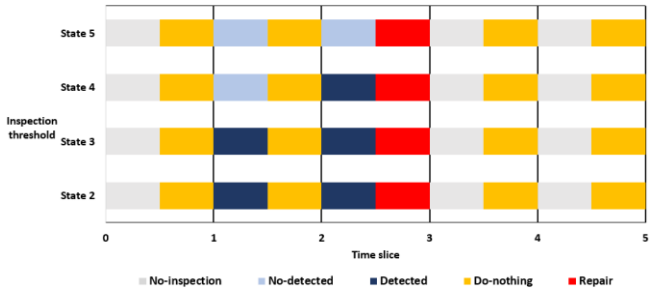
(a) TBM



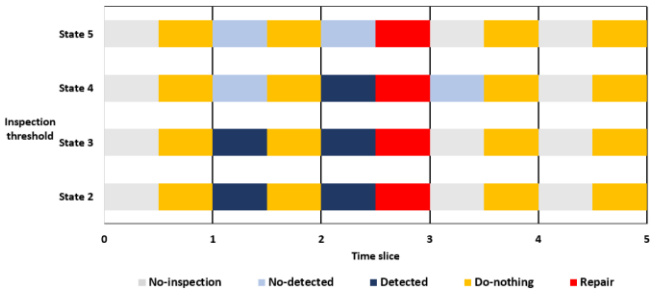
(b) POMDP (R_{MI}=10, R_{FM}=100)



(c) POMDP (R_{MI}=20, R_{FM}=50)



(d) POMDP (R_{MI}=40, R_{FM}=25)



(e) POMDP (R_{MI}=50, R_{FM}=20)

Figure 8. Optimal actions of TBM and POMDP

4. CONCLUSION

In this work, we proposed a maintenance decision-making model for gas turbine components based on POMDP. DBNs and POMDP were integrated to construct the maintenance decision-making model. The fatigue crack growth model was implemented for deterioration of gas turbine engine components, POD curve was used for the inspection model. The total cost of POMDP was lower than that of TBM when inspection quality was high. Also, it was proven that the maintenance decision-making model is more effective than TBM as the cost ratio between maintenance and repair is smaller by parametric study of cost ratio.

Our future work will focus on complicate inspection and maintenance actions. The various options for inspection and maintenance actions will improve the decision-making model more elaborately.

ACKNOWLEDGEMENT

This work was supported by Korea Research Institute for Defense Technology Planning and Advancement (KRIT) Grant funded by Defense Acquisition Program Administration (DAPA) (No. 21-107-E00-012-02).

REFERENCES

Alaswad, S., & Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering and System Safety*, *157*, 54–63. https://doi.org/10.1016/j.ress.2016.08.009

Bousdekis, A., Magoutas, B., Apostolou, D., & Mentzas, G. (2015). A proactive decision making framework for condition-based maintenance. *Industrial Management and Data Systems*, *115*(7), 1225–1250. https://doi.org/10.1108/IMDS-03-2015-0071

C. H. Cook, C. E. Spaeth, D. T. Hunter, & R. J. Hill. (1982, April). Damage Tolerant Design of Turbine Engine Disks. *Turbo Expo: Power for Land, Sea, and Air*. https://doi.org/https://doi.org/10.1115/82-GT-311

Faddoul, R., Raphael, W., Soubra, A.-H., & Chateauneuf, A. (2013). Incorporating Bayesian Networks in Markov Decision Processes. *Journal of Infrastructure Systems*, *19*(4), 415–424. https://doi.org/10.1061/(asce)is.1943-555x.0000134

Hlaing, N., Morato, P. G., Nielsen, J. S., Amirafshari, P., Kolios, A., & Rigo, P. (2022). Inspection and maintenance planning for offshore wind structural components: integrating fatigue failure criteria with Bayesian networks and Markov decision processes. *Structure and Infrastructure Engineering*, *18*(7), 983–1001. https://doi.org/10.1080/15732479.2022.2037667

Lee, B. W., Suh, J., Lee, H., & Kim, T. gu. (2011). Investigations on fretting fatigue in aircraft engine compressor blade. *Engineering Failure Analysis*, *18*(7), 1900–1908. https://doi.org/10.1016/j.engfailanal.2011.07.021

Lee, D., & Achenbach, J. D. (2016). Analysis of the Reliability of a Jet Engine Compressor Rotor Blade Containing a Fatigue Crack. *Journal of Applied Mechanics, Transactions ASME*, *83*(4). https://doi.org/10.1115/1.4032376

Lee, D., & Kwon, K. (2023). Dynamic Bayesian network model for comprehensive risk analysis of fatigue-critical structural details. *Reliability Engineering and System Safety*, *229*. https://doi.org/10.1016/j.ress.2022.108834

Memarzadeh, M., Asce, A. M., Pozzi, M., & Kolter, J. Z. (2014). *Optimal Planning and Learning in Uncertain Environments for the Management of Wind Farms*. https://doi.org/10.1061/(ASCE)CP

Morato, P. G., Nielsen, J. S., Mai, A. Q., & Rigo, P. (2019, May). POMDP based Maintenance Optimization of Offshore Wind Substructures including Monitoring. *ICASP13*. https://doi.org/https://doi.org/10.22725/ICASP13.067

Morato, P. G., Papakonstantinou, K. G., Andriotis, C. P., Nielsen, J. S., & Rigo, P. (2020). *Optimal Inspection and Maintenance Planning for Deteriorating Structural Components through Dynamic Bayesian Networks and Markov Decision Processes*. https://doi.org/10.1016/j.strusafe.2021.102140

Papakonstantinou, K. G., & Shinozuka, M. (2014a). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: POMDP implementation. *Reliability Engineering and System Safety*, *130*, 214–224. https://doi.org/10.1016/j.ress.2014.04.006

Papakonstantinou, K. G., & Shinozuka, M. (2014b). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory. *Reliability Engineering and System Safety*, *130*, 202–213. https://doi.org/10.1016/j.ress.2014.04.005

Maintenance Strategies for Sewer Pipes with Multi-State Degradation and Deep Reinforcement Learning

Lisandro A. Jimenez-Roa [1], Thiago D. Simão [2], Zaharah Bukhsh [2], Tiedo Tinga [1], Hajo Molegraaf [3], Nils Jansen [4,5], and Mariëlle Stoelinga [1,4]

[1] *University of Twente, Enschede, 7522 NB, The Netherlands*
{*l.jimenezroa, t.tinga, m.i.a.stoelinga*}*@utwente.nl*

[2] *Eindhoven University of Technology, Eindhoven, 5612 AE, The Netherlands*
{*t.simao@tue.nl, z.bukhsh*}*@tue.nl*

[3] *Rolsch Assetmanagement, Enschede, 7521 AG, The Netherlands.*
hajo.molegraaf@rolsch.nl

[4] *Radboud University, Nijmegen, 6525 XZ, The Netherlands.*
n.jansen@science.ru.nl

[5] *Ruhr-University Bochum, Bochum, 44801, Germany*

ABSTRACT

Large-scale infrastructure systems are crucial for societal welfare, and their effective management requires strategic forecasting and intervention methods that account for various complexities. Our study addresses two challenges within the Prognostics and Health Management (PHM) framework applied to sewer assets: modeling pipe degradation across severity levels and developing effective maintenance policies. We employ Multi-State Degradation Models (MSDM) to represent the stochastic degradation process in sewer pipes and use Deep Reinforcement Learning (DRL) to devise maintenance strategies. A case study of a Dutch sewer network exemplifies our methodology. Our findings demonstrate the model's effectiveness in generating intelligent, cost-saving maintenance strategies that surpass heuristics. It adapts its management strategy based on the pipe's age, opting for a passive approach for newer pipes and transitioning to active strategies for older ones to prevent failures and reduce costs. This research highlights DRL's potential in optimizing maintenance policies. Future research will aim improve the model by incorporating partial observability, exploring various reinforcement learning algorithms, and extending this methodology to comprehensive infrastructure management.

ABBREVIATIONS

DRL Deep Reinforcement Learning.
IHTMC Inhomogeneous Time Markov Chain.
MDP Markov Decision Process.
MPO Maintenance Policy Optmization.
MSDM Multi-State Degradation Model.
PPO Proximal Policy Optimization.
RL Reinforcement Learning.

1. INTRODUCTION

Sewer network systems, crucial for public health, population well-being, and environmental protection, require maintenance to ensure their reliability and availability (Cardoso et al., 2016). This maintenance is challenged by limited budgets, environmental changes, aging infrastructure, and hard-to-predict system deterioration (Tscheikner-Gratl et al., 2019).

Optimizing maintenance policies for sewer networks requires methodologies that can efficiently explore a broad solution space while adapting to the system's dynamic constraints and complexities. Maintenance Policy Optmization (MPO) addresses these needs by developing and analyzing mathematical models to derive maintenance strategies (De Jonge & Scarf, 2020) that reduce maintenance costs, extend asset life, maximize availability, and ensure workplace safety (Ogunfowora & Najjaran, 2023).

This research explores the potential of Deep Reinforcement Learning (DRL) for MPO of sewer networks, first focusing on a component-level (i.e., pipe-level) analysis. DRL is a framework that merges neural network representation learning capabilities with Reinforcement Learning (RL), a branch of machine learning known for its effectiveness in sequential decision-making problems. RL is increasingly recognized for its role in developing cost-effective policies in MPO across diverse domains such as transportation, manufacturing, civil infrastructure and energy systems. It is emerging as a prominent paradigm in the search for optimal maintenance policies (Marugán, 2023).

This paper aims to achieve two primary objectives: first, to present a comprehensive model for pipe-level MPO analysis facilitated by DRL, considering degradation over the pipe length and employing inhomogeneous-time Markov chain models to simulate the nonlinear stochastic behavior associated with sewer pipe degradation; second, to assess the efficacy of the model's policy through a case study of a large-scale sewer

network in the Netherlands, comparing it with heuristics, including condition-based, scheduled, and reactive maintenance.

We acknowledge as limitations in our approach the focus on *fully observable* state spaces, which means that inspection actions are not necessary, and our analysis is at the *component-level*. Future research will aim to broaden this scope to include partially observable state spaces and system-level analysis.

Contributions. This work's primary contributions include:

(i) We propose a framework to carry out maintenance policy optimization for sewer pipes considering the deterioration along the pipe length. This framework integrates Multi-State Degradation Models (MSDMs) and Deep Reinforcement Learning (DRL).
(ii) Our framework introduces a novel approach by encoding the prediction of the MSDM into the state space, aiming to harness prognostics that describe the degradation pattern of sewer pipes.
(iii) We demonstrate that DRL has the potential to devise intelligent strategic maintenance strategies adaptable to various conditions, such as pipe age.
(iv) We provide our framework in Python and all data used in this study at zenodo.org/records/11258904.

Paper outline. Section 2 presents the technical background. Section 3 outlines our research methodology. Section 4 formulates the MSDM. Section 5 details the framework for maintenance policy optimization via DRL. Section 6 presents our experimental setup. Section 7 analyzes the results. Section 8 discusses findings, concludes, and suggests future research.

Related work. In the past two decades, the need for integral sewer asset management has become evident (Abraham et al., 1998), emphasizing the necessity to understand the mechanisms of deterioration and develop predictive models for proactive and strategic sewer maintenance (Fenner, 2000). Sewer asset management encompasses maintenance, rehabilitation, and inspection and has been investigated through various methodologies, including risk-based strategies (Lee et al., 2021), multi-objective optimization (Elmasry et al., 2019), Markov Decision Processes (Wirahadikusumah & Abraham, 2003), considering the structure of the sewer network (Qasem & Jamil, 2021), machine learning applications (Montserrat et al., 2015; Caradot et al., 2018; Laakso et al., 2019; Hernández et al., 2021), and decision support frameworks (Taillandier et al., 2020; Khurelbaatar et al., 2021; Ramos-Salgado et al., 2022; Assaf & Assaad, 2023).

The integration of RL into sewer asset management is largely unexplored, with existing research mainly concentrating on *real-time control* for smart infrastructure, adapting to environmental changes such as storms. Mullapudi et al. (2020) uses DRL for controlling storm water system valves through simulation of varied storm scenarios. Yin et al. (2023) employ RL for *near real-time* control to minimize sewer overflows. Meanwhile, Zhang et al. (2023) and Tian et al. (2022) both examine improving the robustness of urban drainage systems, the former through decentralized *multi-agent RL* and the latter through *Multi-RL*, with Tian et al. (2024) further improving the model *interpretability* using DRL. Furthermore, Kerkkamp et al. (2022) investigates the sewer network MPO by combining DRL with Graphical Neural Networks to optimize maintenance actions grouping. Jeung et al. (2023) proposes a DRL-based *data assimilation* methodology to enhance storm water and water quality simulation accuracy by integrating observational data with simulation outcomes.

2. TECHNICAL BACKGROUND

2.1. Multi-state degradation model for sewer pipes

The modeling of sewer pipe network degradation has been explored through various methodologies, including physics-based, machine learning, and probabilistic models. For comprehensive discussions on this topic, the reader is directed to Ana & Bauwens (2010); Hawari et al. (2017); Malek Mohammadi et al. (2019); Saddiqi et al. (2023); Zeng et al. (2023).

We adopt a probabilistic approach employing Inhomogeneous Time Markov Chains (IHTMCs) to model the multi-state degradation of sewer pipes. This choice is motivated by the IHTMC's capability to better capture the degradation of long-lived assets such as sewer systems as a non-linear stochastic process, characterized by age-dependent transition probabilities between degradation states (Jimenez-Roa et al., 2024).

Inhomogeneous Time Markov Chains (IHTMCs). An IHTMC is a stochastic process $\{(X_t)\}_{t\geq 0}$, where $t \in [0, \infty)$ is continues and models *time*. The IHTMC is defined as a tuple $M = \langle \Omega, S^0, Q(t) \rangle$, where Ω is a set of K finite states indicating the *state space*, S_k^0 is an *initial-state distribution* on Ω where $\sum_{k\in\Omega} S_k^0 = 1$, and $Q(t) : \Omega \times \Omega \to \mathbb{R}$ is a *time-dependent transition rate matrix*, with entries $q_{ij}(t)$ for $i, j \in \Omega$ and $i \neq j$, representing the rate of transitioning from state i to state j at time t. The diagonal entries $q_{ii}(t)$ are defined such that the sum of each row in $Q(t)$ is zero, ensuring that the *outflow* from any state is equal to the sum of the *inflows* into other states. $Q(t)$ may be parameterized by hazard rates $\lambda(t|\theta)$ derived from the ratio $f(t|\theta)$ and $S(t|\theta)$, being respectively a *probability density function* and a *survival function*, where θ corresponds to the function hyper-parameters. The evolution over time of the IHTMC is governed by the *Forward Kolmogorov* equation:

$$\frac{\partial P_{ij}(t,\tau)}{\partial t} = \sum_{k\in S} P_{ik}(t,\tau) Q_{kj}(t) \tag{1}$$

Here, $P_{ij}(t,\tau) : \Omega \times \Omega \to [0,1]$ is a continuous and differentiable function known as the *transition probability matrix*, indicating the probability of transitioning from state i to state j in the time interval t to τ, where $\tau > t$. From Eq. (1) one can obtain the *master equation of the Markov chain*, which models the flow of probabilities between states by including inflow and outflow terms:

$$\frac{\partial S_k(t)}{\partial t} = \sum_{i\in\Omega, i\neq k} S_i(t) Q_{ik}(t) - S_k(t)\Big(\sum_{j\in\Omega, j\neq k} Q_{kj}(t)\Big) \tag{2}$$

Here, $S_k(t)$ is the probability of being in state $k \in \Omega$ at time t, the term $\sum_{j\in\Omega, j\neq k} Q_{kj}(t)$ represents the rates of transition from state k to all the other states j (excluding self-transitions).

Pipe-element degradation model. We define a pipe element by K sequentially arranged states $S = [S_1, S_2, ..., S_k]$, where S_1 signifies the *pristine* condition and S_k represents the *worst condition*. This categorization is based on sewer network inspection data, which documents types of damage and their severities on a scale from 1 to 5, along with occasional instances of functional failures ($K = 6$). The transitions within our IHTMC, illustrated in Figure 1, permit only progression from a better to a worse state, prohibiting direct improvements without repairs, while allowing any severity level to escalate to functional failure.

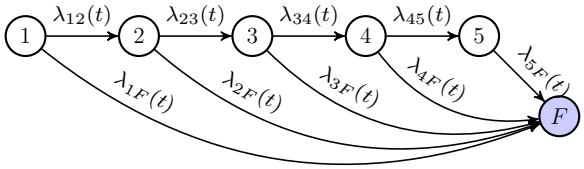


Figure 1. Markov chain structure for IHTMC.

Parametrization of IHTMC. We employed a parameterized approach for IHTMC, involving an assumption on the hazard function. In Section 4.2, we detail the parametrization used in our experimental setup. Several aspects related to the multi-state degradation model, including hyper-parameter tuning and interval-censoring, are beyond the scope of this paper. For further information, we recommend referring to (Jimenez-Roa et al., 2024).

2.2. Markov Decision Process

A Markov Decision Process (MDP) models a stochastic sequential decision process, where both costs and transition functions are dependent solely on the current state and action (Puterman, 1990). Formally, an MDP is described by the tuple $\langle \mathcal{S}, \mathcal{A}, P(s_{t+1}|s_t, a_t), \mathcal{R}(s_t, a_t, s_{t+1}), \pi_0, \gamma \rangle$, with $\mathcal{S}$ as *state space*, $\mathcal{A}$ as the *action space*, $P(s_{t+1}|s_t, a_t)$ as the *transition probability function* indicating the probability of transitioning from state s_t to s_{t+1} given action a_t, where $s_t, s_{t+1} \in \mathcal{S}$ and $a_t \in \mathcal{A}$. The *reward function* $\mathcal{R}(s_t, a_t, s_{t+1})$ specifies the reward for moving from s_t to s_{t+1} by action a_t. The *initial state* π_0 represents the distribution across $\mathcal{S}$, and $\gamma \in [0, 1]$ is the *discount factor* that balances immediate versus future rewards.

2.3. Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) produces virtual agents that interact with environments to learn optimal behaviors through trial and error, as indicated by a reward signal (Arulkumaran et al., 2017). DRL has found applications in robotics, video games, and navigation systems.

We utilize DRL to train agents in virtual environments exhibiting degradation following the MSDM pattern, as detailed in Section 5. Specifically, we apply Proximal Policy Optimization (PPO) (Schulman et al., 2017), a policy gradient method in RL.

PPO aims to optimize the policy an agent uses for action selection, maximizing expected returns. It addresses stability and efficiency issues encountered in previous algorithms like *Trust Region Policy Optimization* by offering a simpler and less computationally expensive method to ensure minor policy updates.

This is achieved through an innovative objective function that penalizes significant deviations from the previous policy, fostering stable and consistent learning. The term "proximal" denotes maintaining proximity between the new and old policies, facilitating a stable training process and rendering PPO popular across various RL applications.

3. Methodology

Our methodology, illustrated in Figure 2, comprises six steps, detailed below.

Step 1. Perform data handling of historical inspection records, selecting subsets (cohorts) of interest, and calibrating the MSDM on this data. This step is beyond the scope of this paper; for details, see Jimenez-Roa et al. (2022, 2024). The results of this step are given in Section 4.

Step 2. After calibrating the MSDM, integrate these models into an environment suitable for RL applications. We present the details of our environment integrating MSDM in Section 5. In addition, we define environments for training RL agents. This is to test different MSDM hypotheses; details on this can be found in Section 6.

Step 3. Train DRL agents with PPO. Use `optuna` for hyperparameter tuning and `Stable Baselines3` for RL implementation. Details are in Section 7.1.

Step 4. Train and select the RL agents with the optimal hyperparameters on the *training* environments. In essence, these agents learn the dynamics described by the MSDM encoded in the environment.

Step 5. Compare the maintenance policies advised by the RL agents using the *test* environment against the heuristics: Condition-Based Maintenance (CBM), Scheduled Maintenance (SchM), and Reactive Maintenance (RM). Find the definition of these heuristics in Section 6.2.

Step 6. Analyze and compare the behavior of the maintenance strategies for the different RL models and heuristics. Reflect on the policies advantages and disadvantages. Find in Section 7.2 the overview of this comparison, and in Section 7.3 are the details along episodes.

4. Multi-state degradation models

4.1. Case study

Our case study conducts a detailed examination of the sewer pipe network in Breda, the Netherlands, which comprises 25,727 sewer pipes covering 1,052 km, mostly built after 1950. The network is primarily made of concrete (72%) and PVC (27%), with the shapes of the pipes being predominantly round (95%) and ovoid (5.4%). These pipes are designed for transportation (98.2%), with 88% being up to 60 meters in length. Additionally, 98.3% have a diameter of up to 1 meter, with the most common diameter being 0.2 meters, and they carry mixed (63%), rain (21%), and waste (16%) contents. The condition of the pipes is evaluated through visual inspections according to the European standard EN 13508 (EN13508, 2012; EN13508-2, 2011), focusing on identifying and classifying damage with specific codes. This study specifically addresses the damage code *BAF*, which signifies *surface damage* and was observed in 35.3% of the inspections.

4.2. Parametrization

We consider three distributions for hazard rate functions: Exponential, Gompertz, and Weibull. The hazard rates $\lambda(t|\cdot)$ for these distributions are specified as follows:

$$\textit{Exponential} \text{ function:} \quad \lambda^E(t|\epsilon) = \epsilon, \tag{3a}$$

$$\textit{Gompertz} \text{ function:} \quad \lambda^G(t|\alpha, \beta) = \alpha\beta e^{\beta t} \tag{3b}$$

$$\textit{Weibull} \text{ function:} \quad \lambda^W(t|\eta, \rho) = \frac{\rho}{\eta}\left(\frac{t}{\eta}\right)^{\rho-1} \tag{3c}$$

In Eq. (3a), a constant hazard rate indicates that the degradation model assumes a *homogeneous* time, exhibiting *memoryless* properties. Eq. (3b) and Eq. (3c) present varying hazard rates, which indicates *inhomogeneous* time.

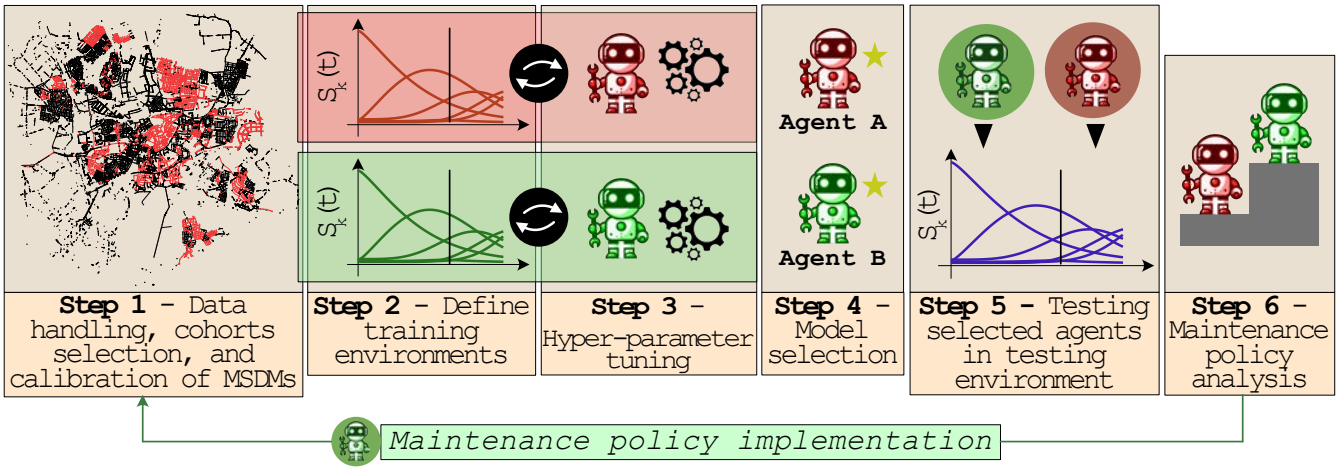


Figure 2. Methodology overview for sewer pipe maintenance policy optimization using Deep Reinforcement Learning and Multi-State Degradation models.

4.3. Solving the Multi-State Degradation Model

In Figure 1, we defined the structure of the Markov chain to model degradation in a sewer pipe, and in Section 4.2 we introduced the hazard rate functions. In the following, we present the corresponding system of differential equations.

$$\frac{\partial S_1(t)}{dt} = -\big(\lambda_{12}(t|\cdot) + \lambda_{1F}(t|\cdot)\big)S_1(t) \tag{4a}$$

$$\frac{\partial S_2(t)}{dt} = \lambda_{12}(t|\cdot)S_1(t) - \big(\lambda_{23}(t|\cdot) + \lambda_{2F}(t|\cdot)\big)S_2(t) \tag{4b}$$

$$\frac{\partial S_3(t)}{dt} = \lambda_{23}(t|\cdot)S_2(t) - \big(\lambda_{34}(t|\cdot) + \lambda_{3F}(t|\cdot)\big)S_3(t) \tag{4c}$$

$$\frac{\partial S_4(t)}{dt} = \lambda_{34}(t|\cdot)S_3(t) + \big(-\lambda_{45}(t|\cdot) - \lambda_{4F}(t|\cdot)\big)S_4(t) \tag{4d}$$

$$\frac{\partial S_5(t)}{dt} = \lambda_{45}(t|\cdot)S_4(t) - \lambda_{5F}(t|\cdot)S_5(t) \tag{4e}$$

$$\frac{\partial S_F(t)}{dt} = \lambda_{1F}(t|\cdot)S_1(t) + \lambda_{2F}(t|\cdot)S_2(t) + \lambda_{3F}(t|\cdot)S_3(t) + \lambda_{4F}(t|\cdot)S_4(t) + \lambda_{5F}(t|\cdot)S_5(t) \tag{4f}$$

Eq. 4 is solved using numerical methods, specifically the `LSODA` algorithm from the FORTRAN `odepack` library implemented in SciPy (Jones et al., 2001–). This algorithm solves systems of ordinary differential equations by employing the `Adams/BDF` method with automatic stiffness detection.

4.4. Parametric Multi-State Degradation Models

We extract a subset from our case study data set to construct a cohort with concrete sewer pipes carrying *mixed and waste content* (cohort `CMW`), representing 37.1% of the sewer network. The model parameters for this cohort are detailed in Appendix A in Tables 7 and 8.

Figure 3 illustrates the MSDMs predictions, detailing the stochastic dynamics of sewer pipe degradation for pipes in cohort `CMW`. As Figure 1 describes, this degradation is segmented into five sequentially ordered severity levels ($k = 1$ to $k = 5$), plus a functional failure state ($k = F$). Differences in the y-axis scales are intentional, to emphasize details and behaviors that various degradation models express across severity levels.

Gray circles represent the frequency per severity level from the inspection dataset. Jimenez-Roa et al. (2022) details how these frequencies are computed. Vertical black lines in Figure 3 mark the last available data point for each severity level.

Additionally, Figure 3 presents the *Turnbull* non-parametric estimator, which assumes no specific distribution for survival times (Turnbull, 1976). In our context, this estimator represents the ground truth of stochastic degradation behavior in sewer pipes.

Tables 1 presents the Root Mean Square Error (RMSE) computed with respect to the Turnbull estimator, for each MSDM assumption, for cohorts `CMW`. These results show that models employing Gompertz and Weibull distributions yield smaller RMSEs compared to the one using the Exponential distribution.

Table 1. RMSE with respect Turnbull estimator, per severity level k and total RMSE, cohort: `CMW`.

	Exponential	Gompertz	Weibull
$S_{k=1}(t)$	3.38E-02	3.27E-02	3.34E-02
$S_{k=2}(t)$	7.04E-02	3.70E-02	3.57E-02
$S_{k=3}(t)$	6.27E-02	2.81E-02	4.38E-02
$S_{k=4}(t)$	4.28E-03	1.13E-02	5.06E-03
$S_{k=5}(t)$	8.33E-03	1.09E-02	3.04E-02
$S_{k=F}(t)$	9.19E-03	1.17E-02	3.62E-03
Total	4.13E-02	2.45E-02	2.96E-02

These MSDMs serve two crucial roles within our environment: first, they drive the degradation behavior of sewer pipes, effectively emulating how sewer pipes degrade over time. Second, the output from the MSDMs is incorporated as prognostic information, available to the agent to support decisions at any

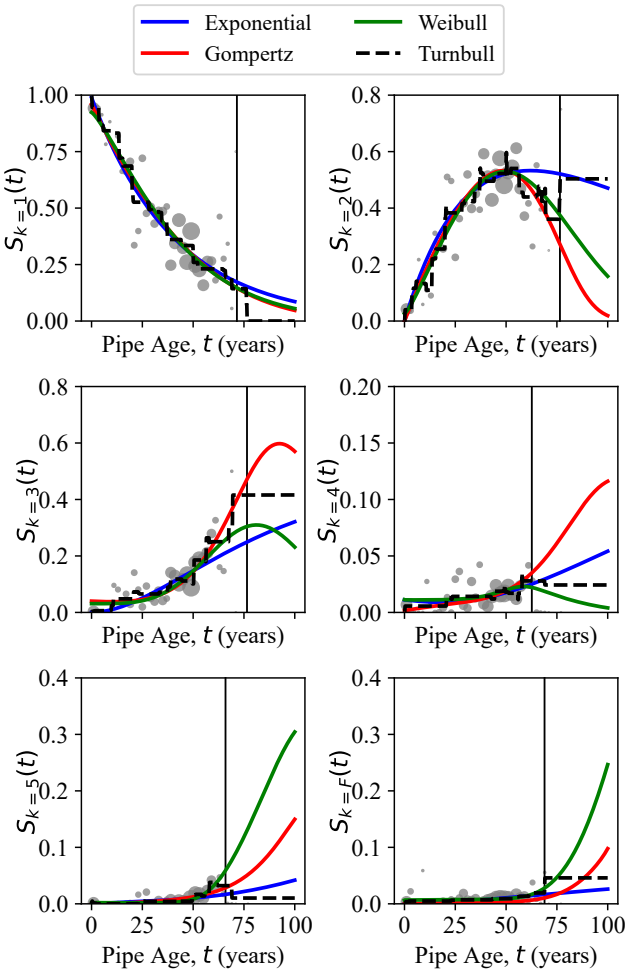


Figure 3. Probability of being in state $k \in \Omega$ at pipe age t $S_k(t)$, using three hazard functions modeled via Exponential, Gompertz, and Weibull probability density functions. The Turnbull non-parametric estimator indicates the ground truth. The gray circles indicate the frequency based on the inspection data set.

time point. This latter aspect is considered a novel feature of our framework. Details on the MDP are provided in the section below.

5. Definition of Markov Decision Process for Maintenance Policy Optimization of a Sewer Pipe considering pipe length degradation

Figure 4 provides the workflow that the RL agent uses to learn maintenance policies for sewer pipes, considering degradation along the pipe length. In the following sections, we provide the details of the environment, namely the state and action spaces, as well as the transition probability and reward functions.

5.1. State space ($\mathcal{S}$)

Our approach focuses on developing age-based maintenance policies, incorporating the sewer pipe's age into the state representation. Our state space is *continuous* and it is structured to include three key components: (i) the age of the pipe, (ii) the *health vector*, and (iii) the stochastic prediction of severity levels. We next describe the last two components.

5.1.1. Health vector (h)

In modeling the degradation of linear structures like sewer pipes, it is essential to represent changes accurately along their length. For this purpose, we define a *health vector* (**h**), which quantitatively measures the degradation at various points along the pipe. The vector is crucial in our framework, particularly influencing the reward function as described in Section 5.4.

Construction of h: We discretize the pipe into segments of equal length ΔL, with $\Delta L < L$, where L is the total length of the pipe. The number of segments, n_d, is calculated using the ceiling function to ensure it remains an integer even if L is not perfectly divisible by ΔL:

$$n_d = \left\lceil \frac{L}{\Delta L} \right\rceil \tag{5}$$

Each segment's degradation level is initially assessed and categorized into *severity levels* according to the MSDM. As the degradation progresses, the state of each segment changes following the transition probabilities described by the matrix $P_{i,j}$, where i is the current severity level, and j is the subsequent severity level, as described by the forward Kolmogorov equation (Eq. 1).

Notice that by doing this, we assume there is no statistical dependency between segments, which is a strong assumption that needs further research. However, for simplicity, we maintain this assumption in our degradation model.

Quantifying Degradation: The distribution of severity levels across the pipe is captured in vector **d**, with each element indicating the severity level of a segment. To quantify this distribution in the health vector **h**, we first count the number of segments at each severity level k using the following expression:

$$n_{d_k} = \sum_{i=1}^{n_d} \mathbf{1}_{\{\mathbf{d}_i = k\}} \tag{6}$$

where $\mathbf{1}$ is the indicator function that is 1 if the condition is true and 0 otherwise. The health vector **h** is then determined by normalizing these counts to reflect the proportion of segments at each severity level:

$$\mathbf{h}_k = \frac{n_{d_k}}{n_d} \tag{7}$$

Here, n_{d_k} is the number of segments at severity level k. Thus, $\mathbf{h}_k$ becomes part of the state space indicating the *level of degradation* present in the pipe.

5.1.2. Stochastic prediction of severity levels

To enable the agent to access information provided by the MSDM, we incorporate the prediction of severity levels into the state space. This is accomplished by solving Eq. 2, yielding a distribution $S_k(t)$.

Finally, our state space is defined as a tuple with 13 elements:

$$\mathcal{S} = \langle \text{Pipe Age}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4, \mathbf{h}_5, \mathbf{h}_F, S_1, S_2, S_3, S_4, S_5, S_F \rangle$$

5.2. Action space ($\mathcal{A}$)

Our action space $\mathcal{A}$ is *discrete* with dimensionality $|\mathcal{A}| = 3$. At each time step t, the agent selects an action a_t. If the decision at time t is *do nothing*, a_t is set to 0. To perform *maintenance*, a_t is set to 1, and to *replace* the pipe, a_t is set to 2. The outcomes of these actions are discussed in Section 5.3.

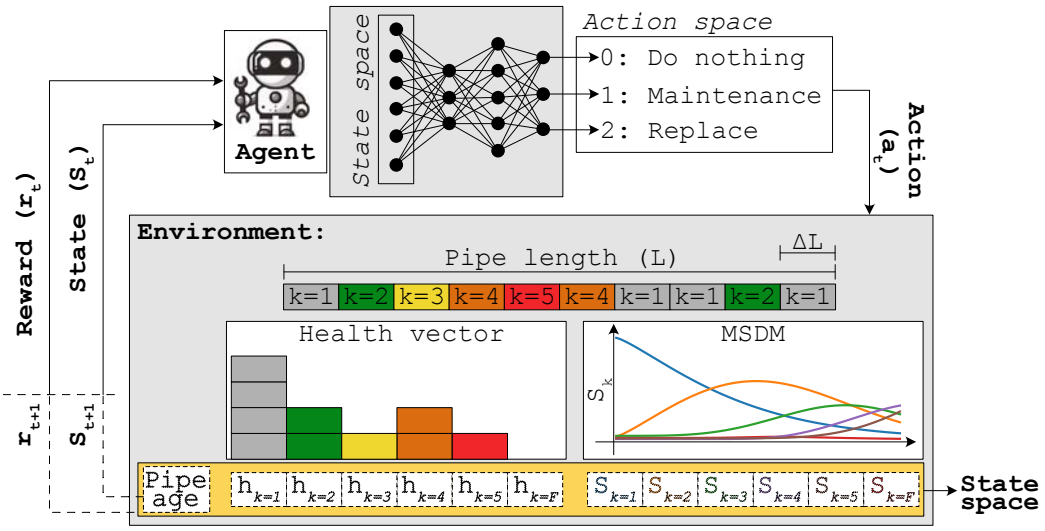


Figure 4. Environment for maintenance policy optimization of a sewer pipe via Deep Reinforcement Learning, considering degradation along the pipe length.

5.3. Transition function (P)

Our transition function $P(s_{t+1}|s_t, a_t)$ is *stochastic*, dependent on time t, and considers both the actions $a \in \mathcal{A}$ and the current s_t and next state s_{t+1} dynamics described by the MSDM. We illustrate the behavior of P with the following example.

For a 30-year-old pipe with length $L = 40$ meters and discretized in segments of length $\Delta L = 1$, let the current state space be $s_{t=30} \in \mathcal{S}$:

$$s_{t=30} = \langle 30, 0.60, 0.35, 0.025, 0.025, 0.0, 0.0, 0.475, 0.436, 0.069, 0.010, 0.005, 0.005 \rangle .$$

$s_{t=30}$ indicates the age of the pipe is 30 years. From Eq. 7, the number of segments at severity k is determined by multiplying the health vector ($\mathbf{h}_k$):

$$\mathbf{h}_k = [0.60, 0.35, 0.025, 0.025, 0.0, 0.0]$$

by 40 meters, yielding $n_{d_k} = [24, 14, 1, 1, 0, 0]$, indicating that, out of the 40 meters of pipe length, 24 segments of 1 meter are at severity $k = 1$, 14 at severity $k = 2$, and so forth.

The distribution $S_k(t = 30.0)$ predicts the probability of being in a severity level k at age $t = 30$. This is achieved by evaluating $t = 30.0$ in the corresponding MSDM.

$$S_k(t = 30.0) = [0.475, 0.436, 0.069, 0.010, 0.005, 0.005]$$

Assuming the agent takes an action every half year, we illustrate the effect of each action in $\mathcal{A}$ below.

- If $a_t = 0$: the agent decides to "do nothing", the pipe's degradation evolves in line with the MSDM progression. Here the new state space becomes $s^{a=0}_{t=30.5}$.

$$s^{a=0}_{t=30.5} = \langle 30.5, 0.575, 0.35, 0.05, 0.025, 0.0, 0.0, 0.470, 0.439, 0.071, 0.010, 0.05, 0.05 \rangle$$

  Notice that the pipe age increased to 30.5, and $n_{d_k} = [23, 14, 2, 1, 0, 0]$, where a segment with severity $k = 1$ progressed to $k = 2$, and one segment with $k = 2$ advanced to $k = 3$. Additionally, $S_k(t)$ is updated by evaluating $t = 30.5$.
- If $a_t = 1$: the agent decides to "perform maintenance," all damage points with severity levels $k \in \{3, 4, 5\}$ are moved to $k = 2$. Consequently, this action does not affect damage points with severity levels $k \in \{1, 2, F\}$. The new state space becomes $s^{a=1}_{t=30.5}$.

$$s^{a=1}_{t=30.5} = \langle 30.5, 0.60, 0.40, 0.0, 0.0, 0.0, 0.0, 0.47, 0.439, 0.071, 0.010, 0.05, 0.05 \rangle$$

  Notice that the pipe age increased to 30.5, and $n_{d_k} = [24, 16, 0, 0, 0, 0]$. However, $S_k(t)$ is updated by evaluating $t = 30.5$, same as when $a_t = 0$.
- If $a_t = 2$: the agent decides to "replace" the pipe, resetting its condition to as good-as-new. The new state space is $s^{a=2}_{t=0.0}$:

$$s^{a=2}_{t=0.0} = \langle 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.986, 0.014, 0.0, 0.0, 0.0, 0.0 \rangle .$$

  The pipe age is reset to 0.0, with $n_{d_k} = [40, 0, 0, 0, 0, 0]$, and $S_k(t)$ is updated for $t = 0.0$.

5.4. Reward function ($\mathcal{R}$)

Our reward function $\mathcal{R}(s_t, a_t, s_{t+1})$ assigns a reward r_t at every decision point t, determined by the current state s_t and action a_t. This function integrates the costs of maintenance (C_M), replacement (C_R), and failures (C_F). $\mathcal{R}$ is *sparse* because it issues a non-zero value only when failures occur or interventions are undertaken.

Maintenance cost C_M is calculated as per Eq. 8, where it combines a variable cost based on severity k with a fixed logistic cost of €500, covering the expenses related to maintenance.

These costs vary with the severity level k, as detailed in Table 2. Note that no maintenance costs are associated with $k = F$ because maintenance cannot be performed on a segment that has already failed. In this case, the agent must replace.

$$C_M = -(\mathbf{h}_k \cdot c_M^k + 500) \tag{8}$$

Table 2. Maintenance costs per severity k per segment (c_M^k)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=F$
$c_M^k =$	0	0	-€500	-€700	-€900	N.A.

Replacement costs (C_R) is computed with Eq. 9:

$$C_R = -(450 + 0.66D + 0.0008D^2)L \tag{9}$$

Here, L and D denote the pipe's length in meters and diameter in millimetres, respectively. C_R is in Euros (€).

The cost of failure, denoted by C_F, entails assigning a substantial penalty when the agent allows a segment of the pipe to achieve a failure state ($k = F$). This penalty cost is established at €-100,000. Our reward function is then:

$$r_t = \frac{C_M + C_R + C_F}{100'000 + 900 \times 40} = \frac{C_M + C_R + C_F}{136'000} \tag{10}$$

where r_t represents the reward obtained at time t, the normalization constant $136'000$ corresponds to the most expensive penalty possible at time t. Thus, r_t is defined within the interval $[-1, 0]$. This reward function aims for the agent to balance maintenance actions with the prevention of undesirable pipe conditions.

6. Experimental Setup

6.1. Setup

We will evaluate our framework with a single pipe of constant length (40 meters) and diameter (200 mm) from the cohort `CMW`, which carries mixed and waste content. Given the constant dimensions, the replacement cost C_R, as defined in Eq. 9, is €24,560. The pipe age, when initializing the episode, is randomly sampled from the uniform distribution $U \sim [0, 50]$, allowing the agent to learn the behavior of pipes within this age range. Additionally, we evaluate the policy in steps of half a year and $\Delta L = 1$ meter.

In the methodology section, we describe the training of two agents: **Agent-E** and **Agent-G**. **Agent-E** is trained in an environment where sewer pipe degradation follows the MSDM parameterised with an *Exponential* probability density function, while **Agent-G** is trained in an environment where degradation follows the MSDM parameterised with a *Gompertz* probability density function.

Both agents are tested in an environment where sewer pipe degradation follows the MSDM parameterized with the *Weibull* probability density function.

During training, each agent follows a specific state space, defined as follows:

$$\mathcal{S}^{\text{Agent-E}}_{\text{Training}} = \langle \text{Pipe Age}, \mathbf{h}_k^E, S_k^E(t) \rangle \tag{11a}$$

$$\mathcal{S}^{\text{Agent-G}}_{\text{Training}} = \langle \text{Pipe Age}, \mathbf{h}_k^G, S_k^G(t) \rangle \tag{11b}$$

Here, $\mathcal{S}$ represents the state space for each agent during training. The subscripts E and G denote the *Exponential* and *Gompertz* probability density functions, respectively. Each agent's objective is to learn an optimal maintenance strategy based on their environment's dynamics.

For testing, both agents are evaluated in the same environment, with the state space defined as follows:

$$\mathcal{S}^{\text{Agent-E}}_{\text{Testing}} = \langle \text{Pipe Age}, \mathbf{h}_k^W, S_k^E(t) \rangle \tag{12a}$$

$$\mathcal{S}^{\text{Agent-G}}_{\text{Testing}} = \langle \text{Pipe Age}, \mathbf{h}_k^W, S_k^G(t) \rangle \tag{12b}$$

In both cases, $S_k^E(t)$ and $S_k^G(t)$ remain consistent with the training phase, reflecting the MSDM predictions. However, the health vector $\mathbf{h}_k$ follows the degradation behavior described by the *Weibull* probability density function, indicated by the subscript W.

6.2. Comparison of maintenance strategies

We compare the RL agent's performance against maintenance policies based on heuristics. For this, we define the following:

- **Condition-Based Maintenance (CBM)**: Maintenance actions are based on the sewer pipe's condition. Specifically, replacement ($a_t = 2$) is performed if `pipe_age` ≥ 70 or $\mathbf{h}_{k=F} \geq 0.0$; maintenance ($a_t = 1$) is conducted if $\mathbf{h}_{k=4} \geq 0.1$ or $\mathbf{h}_{k=5} \geq 0.05$; otherwise, no action ($a_t = 0$) is taken.
- **Scheduled Maintenance (SchM)**: Actions are time-based. Replacement ($a_t = 2$) is executed if $\mathbf{h}_{k=F} \geq 0.0$; maintenance ($a_t = 1$) occurs every 10 years; otherwise, no action ($a_t = 0$) is taken.
- **Reactive Maintenance (RM)**: Replacement is undertaken only upon pipe failure, i.e., replacement ($a_t = 2$) is performed if $\mathbf{h}_{k=F} \geq 0.0$; otherwise, no action ($a_t = 0$) is taken.

Note that CBM and SchM are defined based on plausible values. However, these heuristics can be further calibrated for enhanced performance, which is beyond the scope of this paper.

7. Results

7.1. Implementation and hyper-parameter tuning

Our framework uses `Stable Baselines3` (Raffin et al., 2021), comprising robust implementations of RL algorithms in PyTorch (Ansel et al., 2024). Specifically, we utilize the PPO algorithm. Hyper-parameter optimization is performed using `optuna` (Akiba et al., 2019), a framework dedicated to automating the optimization of hyper-parameters.

The search space encompasses: exponentially-decaying learning rate with a decay rate of 0.05, with an initial learning rate ranging from 10^{-5} to 10^{-2}, discount factor (γ) from 0.8 to 0.9999, entropy coefficient from 0.0001 to 0.01, steps per update (`n_steps`) from 250 to 3000, batch sizes from 16 to 256, activation functions ('tanh', 'relu', 'sigmoid'), policy network architectures ([16, 16], [32, 32], [64, 64], [32, 32, 32]), and training epochs (`n_epochs`) from 5 to 100.

We set up `optuna` to conduct 500 trials, aiming to maximise cumulative reward in 100 episodes. Table 3 details the optimal

hyper-parameters identified. These parameters are used to obtain the results discussed in Sections 7.2 and 7.3, where our agents are trained over a total of 5 million time steps.

Table 3. Optimal hyper-parameters found using `optuna`.

Hyper-parameter	Value
Learning rate	0.0003
Discount factor	0.995
Entropy coefficient	0.008
Steps per update (`n_steps`)	2080
Batch size	104
Activation function	Sigmoid
Policy network architecture	[32, 32, 32]
Training epochs (`n_epochs`)	50

7.2. Policy analysis: overview

This section offers a broad evaluation of the policies, with a detailed analysis over episodes presented in Section 7.3. We compare the agents' performances with the heuristics detailed in Section 6.2 across 100 simulations in the **test** environment (Eq. 12), considering pipe ages of 0, 25, and 50 years, aiming to evaluate policy efficacy concerning degradation over varying pipe ages.

Table 4 presents the *mean policy cost* for Agent-E, Agent-G, CBM, SchM, and RM, highlighting the best and second-best policies in blue and red, with corresponding means and standard deviations from the simulations.

Table 4. Policy cost comparison: Mean and standard deviation (Std.) of costs for Agent-E, Agent-G, CBM, SchM, and RM, evaluated over 100 episodes in the test environment. Costs, in thousands of Euros (€), for pipe ages of 0, 25, and 50 years.

	Pipe age: 0		Pipe age: 25		Pipe age: 50	
Policy	Mean	Std.	Mean	Std.	Mean	Std.
Agent-E	51.3	80.8	116.5	97.7	156.8	121.2
Agent-G	**39.7**	66.2	**78.7**	96.6	**127.1**	128.3
CBM	51.3	107.2	112.3	88.5	**110.7**	86.6
SchM	**42.5**	70.9	**78.9**	96.4	159.8	95.9
RM	48.6	76.6	135.8	86.5	165.7	80.8

From these results, we observe that Agent-G's policy generally outperforms others for pipe ages of 0 and 25 years, securing a second-best position for pipes aged 50 years. It is noted that the cost of all policies increases with pipe age, which aligns with expectations as older pipes require more interventions.

After reviewing the mean policy costs, our focus shifts to the specific actions involved in each policy. Table 5 provides a summary of the actions executed by each policy across simulations for different pipe ages. For new pipes, the SchM policy leads in maintenance activities ($a_t = 1$), with Agent-G following. In terms of replacements ($a_t = 2$), Agent-E is the foremost in implementing this action, with CMB in second place. Both Agent-G and SchM exhibit lower replacement frequencies, explaining the mean policy costs since maintenance actions incur lower expenses compared to the penalties and replacement costs resulting from pipe failures.

For pipes aged 25 years, Agent-G executes more maintenance actions ($a_t = 1$), similar to SchM. Agent-E opts for no maintenance, aligning more with RM's strategy. Although CBM carries out some maintenance actions, replacement actions predominate, indicating a greater tendency to permit pipe failures, which explains the observed differences in mean policy costs.

Table 5. Percentage of actions per policy obtained with Agent-E, Agent-G, CBM, SchM, and RM, evaluated over 100 episodes in the test environment, for different pipe ages.

Pipe age	Action	Agent-E	Agent-G	CBM	SchM	RM
0	$a_t = 0$	99.5	97.51	99.54	94.76	99.61
	$a_t = 1$	0.0	2.21	0.05	4.95	0.00
	$a_t = 2$	0.5	0.28	0.41	0.29	0.39
25	$a_t = 0$	98.81	94.96	98.14	94.56	98.92
	$a_t = 1$	0.00	4.50	0.62	4.94	0.00
	$a_t = 2$	1.19	0.53	1.24	0.50	1.08
50	$a_t = 0$	98.4	94.52	98.05	93.99	98.68
	$a_t = 1$	0.0	4.43	0.67	4.88	0.00
	$a_t = 2$	1.6	1.05	1.28	1.13	1.32

For pipes aged 50 years, CMB offers the most cost-effective policy, with Agent-G's following. CMB conducts fewer maintenance actions and more replacements than Agent-G, accounting for the cost disparity. The policies of Agent-E, RM, and SchM have similar costs. Despite SchM conducting more maintenance, its high number of replacements suggests the maintenance interval requires adjustment. These results indicate that the strategies of CBM, SchM, and RM are less efficient for older pipes due to their higher failure probability.

Regarding the *mean pipe severity level* to assess the impact of various policies on pipe degradation, as shown in Table 6. Our analysis reveals a notable correlation between the average actions per policy, detailed in Table 5, and the mean pipe severity level. Specifically, the Agent-G control strategy tends to maintain pipes within a severity level of $k \in [1, 2, 3]$, whereas the Agent-E, CBM, SchM, and RM policies often result in higher severity levels $k \in [4, 5, F]$, which correlates with increased policy costs.

Table 6. Percentage of severity level per policy obtained with Agent-E, Agent-G, CBM, SchM, and RM, evaluated over 100 episodes in the test environment, for different pipe ages.

Pipe age	Severity	Agent-E	Agent-G	CBM	SchM	RM
0	$k = 1$	59.77	58.75	59.94	59.84	58.88
	$k = 2$	33.27	39.14	32.67	38.05	33.15
	$k = 3$	5.39	1.70	6.00	1.79	6.36
	$k = 4$	1.38	0.28	1.13	0.26	1.30
	$k = 5$	0.18	0.13	0.25	0.04	0.31
	$k = F$	0.01	0.01	0.01	0.01	0.01
25	$k = 1$	50.49	41.72	46.88	39.07	46.62
	$k = 2$	38.96	55.27	43.09	55.55	40.86
	$k = 3$	8.37	2.63	8.48	4.85	9.80
	$k = 4$	1.37	0.29	1.18	0.41	1.51
	$k = 5$	0.78	0.07	0.36	0.10	1.18
	$k = F$	0.02	0.01	0.02	0.01	0.03
50	$k = 1$	57.93	44.65	55.01	40.92	54.36
	$k = 2$	32.58	51.40	36.14	50.46	33.09
	$k = 3$	7.50	3.29	7.20	7.34	9.32
	$k = 4$	1.31	0.39	1.19	0.59	1.64
	$k = 5$	0.65	0.25	0.43	0.67	1.55
	$k = F$	0.03	0.02	0.02	0.03	0.03

To summarize, our findings indicate that the Agent-G's policy, derived using DRL, implements a dynamic management strategy that varies with the pipe's age. This strategy encompasses a more passive approach with new pipes, transitioning to active intervention as the pipes age. This indicates the agent's preference for more frequent maintenance actions rather than allowing pipe failures, which incur higher penalties and replacement costs.

Moreover, Agent-G outperforms Agent-E, illustrating the impact of the degradation model assumption. Specifically, Agent-G's prognostic model used during training aligns more closely with the test environment's degradation pattern than Agent-E's, potentially explaining why Agent-G is better equipped to navigate and understand the degradation pattern. This, in turn, enables it to devise a more effective maintenance policy by leveraging a more accurate degradation model.

7.3. Policy analysis over episode

In Section 7.2, we present an overview of policy performances. This section delves into the details per episode to provide further understanding on these policies. Figures 5, 6, and 7 detail the performance of the Agent-E, Agent-G, CMB, and SchM policies for pipes with ages 0, 25 and 50, respectively. The RM heuristic is excluded from this analysis due to its straightforward approach: allowing the pipe to fail before replacing it.

Figure 5 shows that for a brand new pipe: (a) Agent-G performs maintenance on the pipe at approximately 32 years old; (b) Agent-E opts to replace the pipe when it is around 35 years old, which may be attributed to the presence of elements with higher severity levels in that specific episode; (c) CBM chooses not to act, which results in the least expensive policy in this comparison. However, it is observed that some pipe sections reach severity level $k = 5$ throughout the episode. Not taking any action is deemed risky since progressing to $k = F$ becomes more likely and incurs higher costs; (d) SchM effectively controls severity levels but is more expensive than Agent-G's policy due to more frequent maintenance actions.

Figure 6 shows that for a pipe aged 25: (a) Agent-G exhibits increased activity, indicating more frequent maintenance actions, especially as the pipe ages to 50, shortening the maintenance intervals; (b) Agent-E postpones any action until the pipe fails, at which point it replaces the pipe with a new one, akin to RM; (c) CBM also initiates maintenance around the pipe's 50-year mark. However, degradation escalates from age 60, leading to failure at 66. The inability to manage this increased severity results in significant penalty costs, diminishing the effectiveness of this policy; (d) Similarly, SchM manages severity levels effectively until the pipe reaches approximately 70 years of age, at which point degradation accelerates, resulting in failure at 73.

Figure 7 shows that for a pipe aged 50: (a) Agent-G opts to replace the pipe at age 50, followed by maintenance in the subsequent time step. This decision is likely influenced by parts of the pipe being at severity levels $k \in 3, 4$. Such a scenario is plausible, as new pipes can exhibit high severity levels at a young age due to defects in the material or errors during the construction and installation process. This concept is represented in the MSDM by the initial probability state vector (S_k^0). Additionally, Agent-G recommends maintenance at the interval when the pipe reaches the age of 26 years; (b) Agent-E suggests replacement at approximately 62 years, without recommending further maintenance; (c) CMB advocates for maintenance at about 65 years, followed by replacement at 70 years, in line with heuristics described in Section 6.2; (d) SchM consistently performs maintenance at regular intervals, yet faces significant degradation, culminating in failure around 97 years.

8. Discussion and Conclusions

In this paper, we explore the applications of Prognostics and Health Management (PHM) in sewer pipe asset management. Our study focuses on component-level (i.e., pipe-level) maintenance policy optimization by integrating stochastic multi-state degradation modeling and Deep Reinforcement Learning (DRL). The goal is to assess the effectiveness of DRL in deriving cost-effective maintenance strategies tailored to the specific conditions and requirements of sewer pipes.

A key contribution of our work is the integration of prognostics models with a maintenance policy optimization framework. We utilize a tailored reward function that aligns with damage severity levels, enabling a more complex and realistic maintenance optimization setup.

Our methodology includes a real-world case study from a Dutch sewer network, which provides historical inspection data. Through hyper-parameter tuning and policy analysis, we benchmark our optimized policies against traditional heuristics, including condition-based, scheduled, and reactive maintenance.

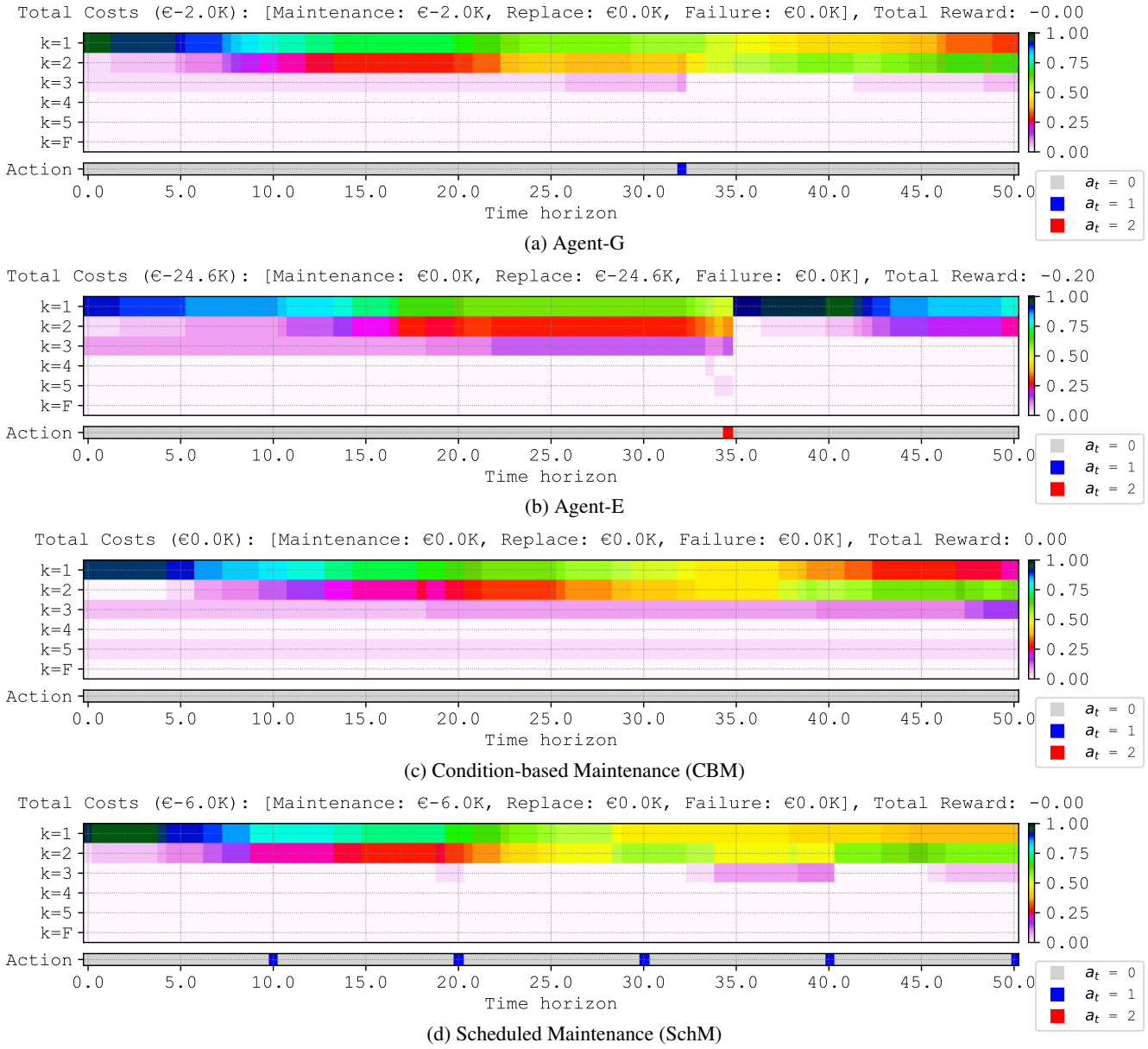
Our findings suggest that agents trained with the Proximal Policy Optimization algorithm are highly capable of developing strategic maintenance policies, adapting to pipe age, and surpassing heuristic baselines by learning cost-effective dynamic management strategies.

To evaluate the impact of degradation model assumptions, we trained one agent using the Gompertz probability density function and another using the Exponential probability density function.

During testing, both agents were assessed in an environment parameterized with the Weibull probability density function. The Gompertz-trained agent, whose behavior more closely resembled the Weibull model, demonstrated better generalization, resulting in more effective maintenance policies compared to the Exponential-trained agent.

Future work: The following directions are identified:

- Advancing toward partially observable state spaces with the introduction of inspection actions, considering context, and leveraging deep learning capabilities.
- Utilizing knowledge acquired by agents to develop explainable and robust heuristics.
- Although this paper focused on a single cohort of pipes, studies in Jimenez-Roa et al. (2022, 2024) show different cohorts exhibit varied dynamics, highlighting the importance of understanding how RL agents adapt.
- Comparing RL-based approaches with other policy optimization algorithms to better understand the capacity of RL methods to achieve global-optima maintenance strategies.
- Investigating various reward functions (e.g., dense) and RL algorithms to determine the most effective for devising maintenance policies.
- Extent to system-level analysis and evaluate aspects such



(a) Agent-G

(b) Agent-E

(c) Condition-based Maintenance (CBM)

(d) Scheduled Maintenance (SchM)

Figure 5. Behavior of policies over an episode for a **new pipe**, showing the health vector over the pipe age and actions per policy: (a) Agent-G, (b) Agent-E, (c) Condition-based Maintenance (CBM), and (d) Scheduled Maintenance (SchM).

as scalability.
- Moving toward multi-infrastructure asset management to promote coordinated management for optimizing costs and minimizing disruption from interventions.

Acknowledgements

This research has been partially funded by NWO under the grant PrimaVera (https://primavera-project.com) number NWA.1160.18.238.

References

Abraham, D. M., Wirahadikusumah, R., Short, T., & Shahbahrami, S. (1998). Optimization modeling for sewer network management. *Journal of construction engineering and management*, *124*(5), 402–410.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.

Ana, E., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods. *Urban Water Journal*, *7*(1), 47–59.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., … others (2024). Pytorch 2: Faster machine

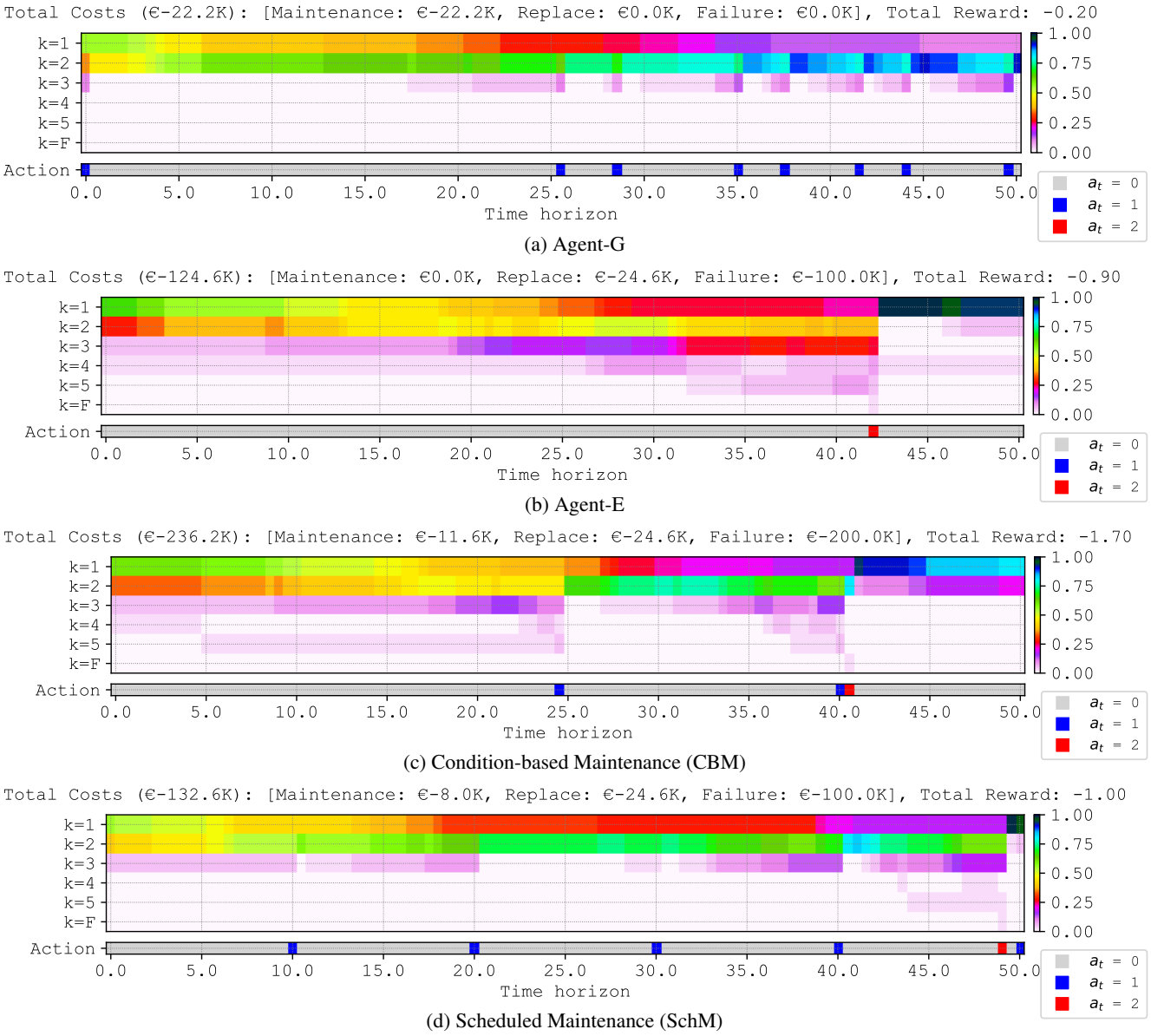


Figure 6. Behavior of policies over an episode for a **pipe aged 25**, showing the health vector over the pipe age and actions per policy: (a) Agent-G, (b) Agent-E, (c) Condition-based Maintenance (CBM), and (d) Scheduled Maintenance (SchM).

learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th acm international conference on architectural support for programming languages and operating systems, volume 2* (pp. 929–947).

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, *34*(6), 26–38.

Assaf, G., & Assaad, R. H. (2023). Optimal preventive maintenance, repair, and replacement program for catch basins to reduce urban flooding: Integrating agent-based modeling and monte carlo simulation. *Sustainability*, *15*(11), 8527.

Caradot, N., Riechel, M., Fesneau, M., Hernandez, N., Torres, A., Sonnenberg, H., … Rouault, P. (2018). Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in berlin, germany. *Journal of Hydroinformatics*, *20*(5), 1131–1147.

Cardoso, M., Almeida, M. d. C., & Santos Silva, M. (2016). Sewer asset management planning–implementation of a structured approach in wastewater utilities. *Urban Water Journal*, *13*(1), 15–27.

De Jonge, B., & Scarf, P. A. (2020). A review on maintenance optimization. *European journal of operational research*, *285*(3), 805–824.

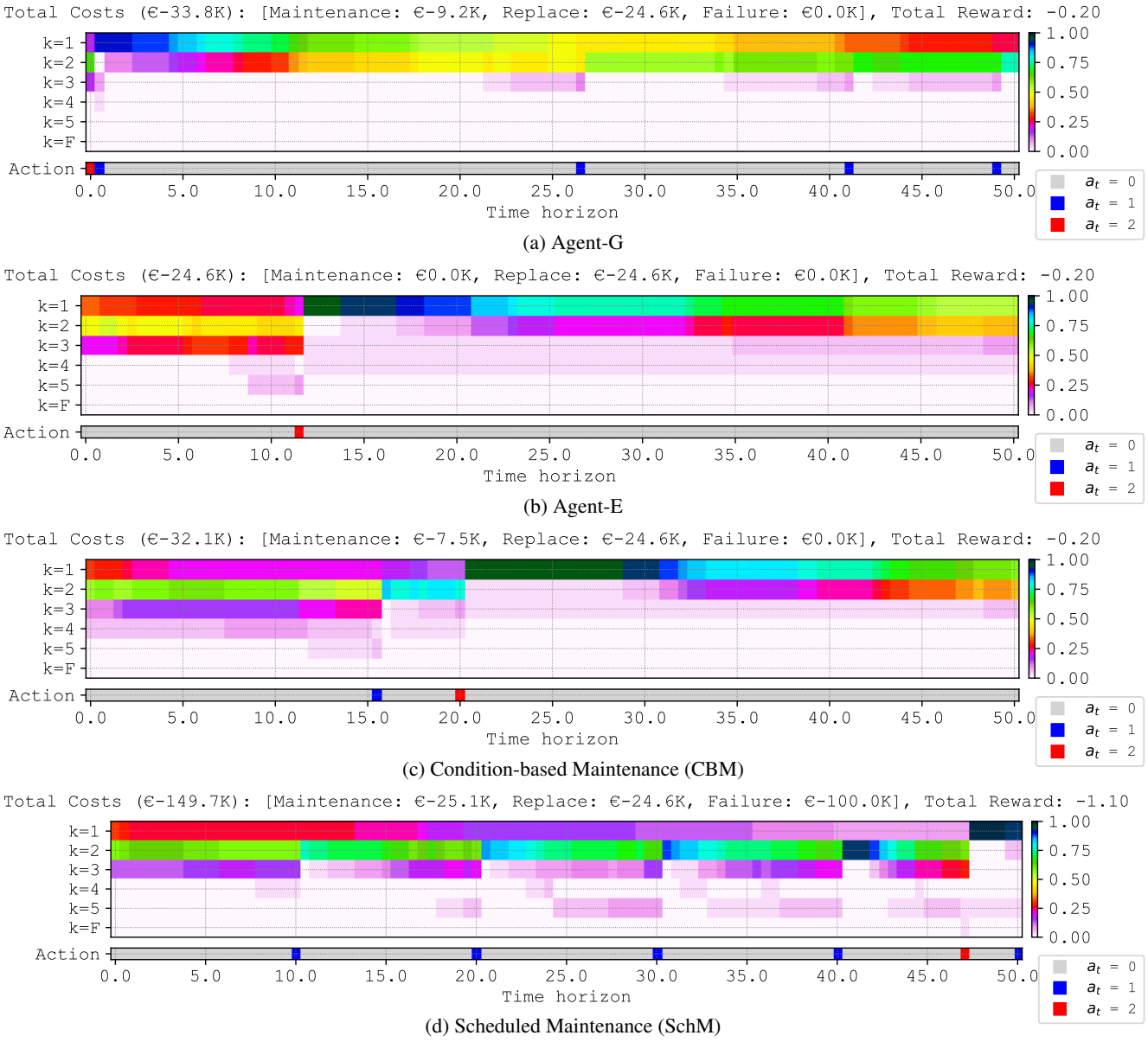


Figure 7. Behavior of policies over an episode for a **pipe aged 50**, showing the health vector over the pipe age and actions per policy: (a) Agent-G, (b) Agent-E, (c) Condition-based Maintenance (CBM), and (d) Scheduled Maintenance (SchM).

Elmasry, M., Zayed, T., & Hawari, A. (2019). Multi-objective optimization model for inspection scheduling of sewer pipelines. *Journal of Construction Engineering and Management*, *145*(2), 04018129.

Fenner, R. A. (2000). Approaches to sewer maintenance: a review. *Urban water*, *2*(4), 343–356.

Hawari, A., Alkadour, F., Elmasry, M., & Zayed, T. (2017). Simulation-based condition assessment model for sewer pipelines. *Journal of Performance of Constructed Facilities*, *31*(1), 04016066.

Hernández, N., Caradot, N., Sonnenberg, H., Rouault, P., & Torres, A. (2021). Optimizing svm models as predicting tools for sewer pipes conditions in the two main cities in colombia for different sewer asset management purposes. *Structure and Infrastructure Engineering*, *17*(2), 156–169.

Investigation and assessment of drain and sewer systems outside buildings - Part 1: General Requirements (Standard). (2012, October). Avenue Marnix 17, B-1000 Brussels: European Committee for Standardization (CEN).

Investigation and assessment of drain and sewer systems outside buildings - Part 2: Visual inspection coding system (Standard). (2011, May). Avenue Marnix 17, B-1000 Brussels: European Committee for Standardization (CEN).

Jeung, M., Jang, J., Yoon, K., & Baek, S.-S. (2023). Data

assimilation for urban stormwater and water quality simulations using deep reinforcement learning. *Journal of Hydrology*, *624*, 129973.

Jimenez-Roa, L. A., Heskes, T., Tinga, T., Molegraaf, H. J., & Stoelinga, M. (2022). Deterioration modeling of sewer pipes via discrete-time markov chains: A large-scale case study in the netherlands. In *32nd european safety and reliability conference, esrel 2022: Understanding and managing risk and reliability for a sustainable future* (pp. 1299–1306).

Jimenez-Roa, L. A., Tinga, T., Heskes, T., & Stoelinga, M. (2024). Comparing homogeneous and inhomogeneous time markov chains for modelling degradation in sewer pipe networks. In *Proceedings of the european safety and reliability conference (esrel 2024).* (Under review)

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). *SciPy: Open source scientific tools for Python.* Retrieved from `http://www.scipy.org/`

Kerkkamp, D., Bukhsh, Z. A., Zhang, Y., & Jansen, N. (2022). Grouping of maintenance actions with deep reinforcement learning and graph convolutional networks. In *Icaart (2)* (pp. 574–585).

Khurelbaatar, G., Al Marzuqi, B., Van Afferden, M., Müller, R. A., & Friesen, J. (2021). Data reduced method for cost comparison of wastewater management scenarios–case study for two settlements in jordan and oman. *Frontiers in Environmental Science*, *9*, 626634.

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2019). Sewer life span prediction: Comparison of methods and assessment of the sample impact on the results. *Water*, *11*(12), 2657.

Lee, J., Park, C. Y., Baek, S., Han, S. H., & Yun, S. (2021). Risk-based prioritization of sewer pipe inspection from infrastructure asset management perspective. *Sustainability*, *13*(13), 7213.

Malek Mohammadi, M., Najafi, M., Kaushal, V., Serajiantehrani, R., Salehabadi, N., & Ashoori, T. (2019). Sewer pipes condition prediction models: A state-of-the-art review. *Infrastructures*, *4*(4), 64.

Marugán, A. P. (2023). Applications of reinforcement learning for maintenance of engineering systems: A review. *Advances in Engineering Software*, *183*, 103487.

Montserrat, A., Bosch, L., Kiser, M., Poch, M., & Corominas, L. (2015). Using data from monitoring combined sewer overflows to assess, improve, and maintain combined sewer systems. *Science of the Total Environment*, *505*, 1053–1061.

Mullapudi, A., Lewis, M. J., Gruden, C. L., & Kerkez, B. (2020). Deep reinforcement learning for the real time control of stormwater systems. *Advances in water resources*, *140*, 103600.

Ogunfowora, O., & Najjaran, H. (2023). Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization. *Journal of Manufacturing Systems*, *70*, 244–263.

Puterman, M. L. (1990). Markov decision processes. *Handbooks in operations research and management science*, *2*, 331–434.

Qasem, A., & Jamil, R. (2021). Gis-based financial analysis model for integrated maintenance and rehabilitation of underground pipe networks. *Journal of Performance of Constructed Facilities*, *35*(5), 04021046.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *The Journal of Machine Learning Research*, *22*(1), 12348–12355.

Ramos-Salgado, C., Muñuzuri, J., Aparicio-Ruiz, P., & Onieva, L. (2022). A comprehensive framework to efficiently plan short and long-term investments in water supply and sewer networks. *Reliability Engineering & System Safety*, *219*, 108248.

Saddiqi, M. M., Zhao, W., Cotterill, S., & Dereli, R. K. (2023). Smart management of combined sewer overflows: From an ancient technology to artificial intelligence. *Wiley Interdisciplinary Reviews: Water*, *10*(3), e1635.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Taillandier, F., Elachachi, S., & Bennabi, A. (2020). A decision-support framework to manage a sewer system considering uncertainties. *Urban Water Journal*, *17*(4), 344–355.

Tian, W., Fu, G., Xin, K., Zhang, Z., & Liao, Z. (2024). Improving the interpretability of deep reinforcement learning in urban drainage system operation. *Water Research*, *249*, 120912.

Tian, W., Liao, Z., Zhi, G., Zhang, Z., & Wang, X. (2022). Combined sewer overflow and flooding mitigation through a reliable real-time control based on multi-reinforcement learning and model predictive control. *Water Resources Research*, *58*(7), e2021WR030703.

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., ... others (2019). Sewer asset management–state of the art and research needs. *Urban Water Journal*, *16*(9), 662–675.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(3), 290–295.

Wirahadikusumah, R., & Abraham, D. M. (2003). Application of dynamic programming and simulation for sewer management. *Engineering, Construction and Architectural Management*, *10*(3), 193–208.

Yin, Z., Leon, A. S., Sharifi, A., & Amini, M. H. (2023). Optimal control of combined sewer systems to minimize sewer overflows by using reinforcement learning. In *World*

environmental and water resources congress 2023 (pp. 711–722).

Zeng, X., Wang, Z., Wang, H., Zhu, S., & Chen, S. (2023). Progress in drainage pipeline condition assessment and deterioration prediction models. *Sustainability*, *15*(4), 3849.

Zhang, Z., Tian, W., & Liao, Z. (2023). Towards coordinated and robust real-time control: A decentralized approach for combined sewer overflow and urban flooding reduction based on multi-agent reinforcement learning. *Water Research*, *229*, 119498.

BIOGRAPHIES

Lisandro A. Jimenez-Roa is a doctoral candidate in Computer Science at the University of Twente, The Netherlands. He has a background in civil engineering and has contributed to various projects in structural health monitoring, finite element modeling, and damage detection through data analytics and machine learning. His current research focuses on Prognostics and Health Management, specifically engineering systems within the PrimaVera project (**https://primavera-project.com**), emphasizing multi-state stochastic degradation modeling and maintenance policy optimization using Reinforcement Learning techniques.

Thiago D. Simão is an Assistant Professor in the Eindhoven University of Technology, the Netherlands. He obtained his PhD in Computer Science at Delft University of Technology. Previously, he was a PostDoc researcher at Radboud University Nijmegen. His research interests lie primarily in reliably automating sequential decision-making, focusing on reinforcement learning.

Zaharah Bukhsh is an assistant professor at Eindhoven University of Technology, Eindhoven, Netherlands . She holds a Master's degree in computer science and a Ph.D. in engineering technology from University of Twente, Enschede, Netherlands. Her research focuses on developing data-driven methods with deep learning and deep reinforcement learning. Her research targets broad application areas including asset management, scheduling, and resource optimization. She has contributed to several H2020 and NWO research projects.

Tiedo Tinga is a full professor in dynamics based maintenance at the University of Twente since 2012 and full professor of Life Cycle Management at the Netherlands Defence Academy since 2016. He received his Ph.D. degree in mechanics of materials from Eindhoven University in 2009. He is chairing the smart maintenance knowledge center and leads a number of research projects on developing predictive maintenance concepts, mainly based on the physics of failure models, but also following data-driven approaches.

Hajo Molegraaf completed his PhD at the University of Groningen and has worked as an assistant and postdoc researcher at the University of Geneva and Yale University. Since October 2022, Molegraaf joined as a Research Fellow within the Formal Methods and Tools (FMT) group in the EEMCS faculty at Twente. Additionally, Molegraaf is a co-founder and software developer at Rolsch Assetmanagement, a company based in Enschede, The Netherlands.

Nils Jansen is a full professor at the Ruhr-University Bochum, Germany, and leads the chair of Artificial Intelligence and Formal Methods. The mission of his chair is to increase the trustworthiness of Artificial Intelligence (AI). Prof. Jansen is also an associate professor at Radboud University, Nijmegen, The Netherlands. He was a research associate at the University of Texas at Austin and received his Ph.D. with distinction from RWTH Aachen University, Germany. His research is on intelligent decision-making under uncertainty, focusing on formal reasoning about the safety and dependability of artificial intelligence (AI). He holds several grants in academic and industrial settings, including an ERC starting grant titled Data-Driven Verification and Learning Under Uncertainty (DEUCE).

Mariëlle Stoelinga is a full professor of risk analysis for high-tech systems, both at the University of Twente and Radboud University, the Netherlands. She holds a Master's degree in Mathematics & Computer Science, and a Ph.D. in Computer Science. After her Ph.D., she has been a postdoctoral researcher at the University of California at Santa Cruz, USA. Prof. Stoelinga leads various research projects, including a large national consortium on Predictive Maintenance and an ERC consolidator grant on safety and security interactions.

APPENDIX A. PARAMETERS OF MULTI-STATE DEGRADATION MODELS

Table 7. MSDM hyper-parameters for cohort `CMW`, using hazard functions modeled with the *exponential* ($\lambda^E(t|\epsilon)$), *Gompertz* ($\lambda^G(t|\alpha,\beta)$), and *Weibull* ($\lambda^W(t|\eta,\rho)$) probability density functions.

	$\lambda^E(t\|\epsilon)$	$\lambda^G(t\|\alpha,\beta)$		$\lambda^W(t\|\eta,\rho)$	
$i \to j$	ϵ	α	β	η	ρ
$1 \to 2$	2.4E-02	2.3E+00	8.4E-03	1.3E+00	4.4E+01
$2 \to 3$	9.4E-03	2.1E-02	5.5E-02	2.9E+00	7.7E+01
$3 \to 4$	5.7E-03	3.3E+00	2.8E-03	3.5E+00	8.1E+01
$4 \to 5$	1.8E-02	2.4E+00	8.7E-03	7.0E+00	5.5E+01
$1 \to F$	3.0E-18	1.4E-01	3.1E-04	4.1E-06	4.6E+01
$2 \to F$	6.0E-04	8.8E-01	7.0E-19	2.7E-04	4.6E+01
$3 \to F$	1.0E-18	2.2E-03	4.5E-02	3.0E-05	4.7E+01
$4 \to F$	1.0E-18	9.8E-05	8.6E-03	1.1E-03	4.5E+01
$5 \to F$	1.0E-18	7.0E-19	3.8E-01	1.7E+00	5.9E+01

Table 8. Initial state vector S_k^0 for MSDM of cohort `CMW`.

S_k^0	Exponential	Gompertz	Weibull
$k=1$	9.89E-01	9.58E-01	9.23E-01
$k=2$	1.26E-17	0.00E+00	2.59E-02
$k=3$	3.70E-23	4.00E-02	3.10E-02
$k=4$	1.11E-02	1.61E-03	1.13E-02
$k=5$	2.11E-22	2.00E-15	2.07E-03
$k=F$	3.87E-22	1.56E-04	6.40E-03

Model-based Probabilistic Diagnosis in Large Cyberphysical Systems

Giso Dal[1], Arjen Hommersom[2], Guus Grievink[3], and Peter J.F. Lucas[4]

[1,3,4] *EEMCS Faculty, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*
gisodal@gmail.com, g.grievink@student.utwente.nl, peter.lucas@utwente.nl

[2] *Open University, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands*
arjen.hommersom@ou.nl

ABSTRACT

Model-based diagnosis is concerned with diagnosing faults or malfunction of real-world physical or cyberphysical systems using a model of the structure and behavior of the systems. As cyberphysical systems can be extremely large and complex, and the associated computational models will be then equally large and complex, they impose a hard to beat challenge on the computational feasibility of reasoning with such models. When such a model is able to handle the uncertainty associated with diagnostics, giving rise to probabilistic model-based diagnostics, the computational feasibility becomes even harder. This paper: (1) proposes a novel graphical method underlying model-based diagnostics; (2) demonstrates experimentally how a novel, by the authors developed architecture of partitioned positive weighted model counting, is able to handle exact inference to answer a variety of probabilistic queries regarding the health status of a cyberphysical system. Results obtained are well within acceptable time bounds.

1. INTRODUCTION

Cyberphysical systems combine and integrate physical and computational processes often with a special role for sensor information (Lee, 2008). Nowadays, because of the explosive rise in the role of embedded software in physical systems, there are many of such systems, for example industrial printing or vending machines. In particular in a commercial setting, such machines are desired to experience the least possible downtime, as being out of order usually has undesirable, often financial, consequences for the users. Thus there is a need to find the causes of malfunction as quickly as possible, a process usually referred to a *diagnosis*, a terms taken from the field of medicine (Lucas, 1996); the terms *troubleshooting* and *fault-finding* are also often used in engineering. In the case of cyberphysical systems faults or defects concern physical or software components, or possibly their interaction. The purpose of the diagnostic process consists of automated fault finding followed by repair or assisting technicians in a repair job on site (Grievink, 2022).

Automated computer-based diagnosis in engineering has a long-standing tradition, where in particular fault tree analysis is a commonly used technique (Ruijters & Stoelinga, 2015). However, other automated fault detection and analysis methods have also been developed (Dowdeswell, Sinha, & MacDonell, 2020). In the present paper we depart from a framework developed in the 1980s by Johan de Kleer, and which is known as *model-based diagnosis*, MBD for short (de Kleer & Williams, 1987). The adjective 'model-based' comes from the principle that with the design of a machine one possesses already valuable knowledge about its structure or architecture and its functional components and their interactions before the machine is actually produced, purchased, and employed in practice. This knowledge can be put to use in a diagnostic setting.

De Kleer's method of model-based diagnosis is based on comparing qualitative predictions of behavior of a *model* of a given machine with actual observations, which explains why it is also called *consistency-based diagnosis* (CBD) (Reiter, 1987). CBD is traditionally seen as a kind of symbolic or logical assumption-based reasoning (Genesereth & Nilsson, 1987). However, even in the early days of MBD it was realized that probabilistic information could play a role in improving the accuracy of diagnostic solutions (de Kleer, 1991). It was subsequently proved by Judea Pearl that MBD can be mapped to the multivariate probabilistic representation of Bayesian networks (Pearl, 1988). The advantage of Bayesian networks as a formalism of model-based diagnosis is that in principle uncertain, probabilistic knowledge about the occurrence of faults can be integrated and also learned from data.

However, a well-known problem of MBD, and probabilistic MBD is no exception, is that models can be very large and thus inference is often infeasible. There can be little doubt that with the large and complex cyberphysical systems developed today, and even more so in the future, building and performing inference with such models will hit computational obstacles. During the last few years we have been working on moving the boundaries of probabilistic inference by developing a novel framework referred to as *partitioned positive weighted model counting* (that can also be parallelized), inspired by the success of model counting in software verification (Dal & Lucas, 2017; Dal, Laarman, Hommersom, & Lucas, 2021). This framework was shown in various papers to be superior to other probabilistic methods (Dal et al., 2021; Dal, Laarman, & Lucas, 2023). This made us wonder whether it might be a good candidate for model-based diagnosis of large cyberphysical systems. Positive weighted model counting exploits symmetries in probability tables, which typically also occur in models used in MBD.

The main contributions of this paper are as follows:

- A new representation of compositional Bayesian networks that supports developing large probabilistic model-based systems from specifications of system components;
- A novel method of probabilistic model-based diagnosis, we call it *Bayesian* model-based diagnosis, that properly takes into account the dependences between components in MBD, different from an earlier developed and limited method (Grievink, 2022);
- Experimental evidence that our publicly available software tool PARAGNOSIS [1] (Dal et al., 2023), implementing partitioned weighted model counting, supports solving diagnostic problems of systems with different size and complexity, including very large and complex ones.

The organization of this paper is as follows. First, in Section 2, some basic principles of consistency-based diagnosis are introduced, follows by a compact summary of the method of weighted model counting, and the mapping of MBD to Bayesian networks. As our work on partitioned positive weighted model counting has been extensive published, we refer for details about how it works to those publications (Dal & Lucas, 2017; Dal, Michels, & Lucas, 2017; Dal et al., 2021, 2023). In Section 3 the compositional method of assumption-based Bayesian model-based diagnosis is developed. Experimental results are summarized in Section 4, which is followed by conclusions and a discussion of the results in Section 5.

2. BACKGROUND

2.1. Consistency-based Diagnosis

Given specific input and output of a cyberphysical system, the output of the model of the system is compared with the observed output of the (real) system. A discrepancy between these two indicates a fault or malfunction in the actual system and explains the name 'consistency-based diagnosis' (Reiter, 1987). Below, some of the common definitions that occur in the literature on CBD are repeated and adapted, where in particular (de Kleer, Mackworth, & Reiter, 1992) is followed.

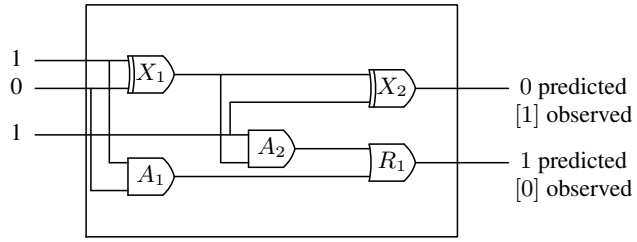


Figure 1. Full adder with inputs and observed and predicted outputs. Here, Obs $= \{\text{in1}(X1) = 1, \text{in2}(X1) = 0, \text{in1}(A2) = 1, \text{out}(X2) = 1, \text{out}(R1) = 0\}$.

A *diagnostic problem* DP is defined as a system SYS together with a *set of observation* Obs: DP $= (\text{SYS}, \text{Obs})$. A *system* SYS consists of a *system description* SD and a set of *components* Comps: SYS $= (\text{SD}, \text{Comps})$. The system description defines the normal behavior of components and how these components are connected by means of first-order logical sentences. Given SYS, a *diagnosis* $D \subseteq$ Comps is defined as a *subset-minimal* set of components that, when behaving abnormally, explains the observation of a faulty system. Formally, a diagnosis is defined as a subset-minimal set $D \subseteq$ Comps such that:

$$\text{SD} \cup \text{Obs} \cup \{\text{Ab}(c) \mid c \in D\} \cup \{\neg\text{Ab}(c) \mid c \in \text{Comps} \backslash D\} \not\models \bot \quad (1)$$

where 'Ab' is the abnormality predicate that indicates that a component c behaves abnormally (and thus $\neg$Ab indicates normal behavior) and $\bot$ is falsum (the left-hand side of $\not\models$ is consistent).

Example 2.1 (Adapted from (Reiter, 1987))**.** *Consider the logical circuit depicted in Fig. 1, which represents a full adder, i.e. a circuit that can be used for the addition of two bits with carry-in and carry-out bits. This circuit consists of two* AND *gates (*$A1$ *and* $A2$*), one OR gate (*$R1$*) and two exclusive-*OR *(*XOR*) gates (*$X1$ *and* $X2$*);* $\text{Comps} = \{A1, A2, X1, X2, R1\}$. *The input and output of components* c *are denoted as* $\text{in}(c)$ *and* $\text{out}(c)$*, respectively.*

The behavior description SD *consists of the following axioms:*

$$\neg\text{Ab}(c) \rightarrow \text{out}(c) = \text{and}(\text{in1}(c), \text{in2}(c)), \textit{ for } c \in \{A1, A2\},$$
$$\neg\text{Ab}(c) \rightarrow \text{out}(c) = \text{xor}(\text{in1}(c), \text{in2}(c)), \textit{ for } c \in \{X1, X2\},$$
$$\neg\text{Ab}(c) \rightarrow \text{out}(c) = \text{or}(\text{in1}(c), \text{in2}(c)), \textit{ for } c = R1.$$

These logical rules describe the normal behavior of each individual component (gate).

[1] https://github.com/gisodal/paragnosis

The component connections are described as follows:

$$\begin{aligned} \text{out}(X1) &= \text{in2}(A2) & \text{out}(X1) &= \text{in1}(X2) \\ \text{out}(A2) &= \text{in1}(R1) & \text{in1}(A2) &= \text{in2}(X2) \\ \text{in1}(X1) &= \text{in1}(A1) & \text{in2}(X1) &= \text{in2}(A1)\ . \\ \text{out}(A1) &= \text{in2}(R1) \end{aligned}$$

With the observations Obs *as indicated in Fig. 1 it is clear that when assuming the empty diagnosis, $D = \emptyset$ — all components are behaving normally — 1 will give an inconsistency, as also indicated in the figure (predicted and observed outputs differ). There are multiple solutions for the diagnostic problem in this case. For example, $D = \{X1\}$, $D' = \{X2, R1\}$, and $D'' = \{X2, A2\}$ are diagnoses.*

2.2. Weighted Model Counting

2.2.1. Bayesian Networks

A *Bayesian network* (BN) $\mathcal{B} = (G, P)$ is a directed acyclic graph $G = (V, A)$ that associates 1–1 random variables X_v to nodes $v \in V$ in the graph (Pearl, 1988). The directed edges $(v, w) \in A$ represent conditional (in)dependence assumptions and P stands for a joint probability distribution of the set of variables X_V defined as follows:

$$P(X_V = x_V) = \prod_{v \in V} P(X_v = x_v \mid X_{\pi(v)} = x_{\pi(v)}) \quad (2)$$

Thus, a BN is defined in terms of a (family of) conditional probability distributions of $X_v \in X_V$ given the variables corresponding to the parents $\pi(v)$ of $v \in V$ in the graph, i.e., $X_{\pi(v)}$, specified as $P(X_v \mid X_{\pi(v)})$, called *conditional probability tables* or CPTs for short in the following.

Posterior probability distributions of the form

$$P(X_U \mid \text{Evidence}), \quad (3)$$

with 'Evidence' a set of *observations* or measurements concerning particular variables $X_v \in X_V$, with typically $v \notin U$, $U \subseteq V$, can be computed based on the specification of a BN —a process called *probabilistic inference* or *reasoning*— using common axioms of probability theory. However, for real-life networks advanced algorithms are required as the computation is NP-hard in general and often quite intensive for real-life networks (Koller & Friedman, 2009).

By exploiting the conditional independence assumptions, BNs represent concise factorizations of a joint probability distribution. The size of the factorization has direct implications toward the cost of probabilistic inference. A more expressive model must be used to in order to exploit properties of CPTs (Chavira & Darwiche, 2008). A prominent way of achieving this is to find a more concise and canonical representation such as a Binary Decision Diagram (BDD) (Bryant, 1986). Compiling a BN to a decision diagram (DD) representation is commonly referred to as *knowledge compilation* (Darwiche & Marquis, 2002), or simply compilation.

Table 1. Three examples of models of the encoding of variable X and associated probability distribution.

	Models					Associated probability
1	x_1	$\overline{x_2}$	$\overline{x_3}$	ω_1	$\overline{\omega_2}$	$W(\omega_1)W(\overline{\omega_2})=0.8 \cdot 1 = 0.8$
2	$\overline{x_1}$	x_2	$\overline{x_3}$	$\overline{\omega_1}$	ω_2	$W(\overline{\omega_1})W(\omega_2)=1 \cdot 0.1 = 0.1$
3	$\overline{x_1}$	$\overline{x_2}$	x_3	$\overline{\omega_1}$	ω_2	$W(\overline{\omega_1})W(\omega_2)=1 \cdot 0.1 = 0.1$

2.2.2. Encoding

Prior to compiling a BN to a DD, we require an encoding to transition from the multi-valued domain of discrete random variables to the Boolean domain. There are multiple ways to do this. We choose to first translate a BN to a Boolean formula with dedicated variables to represent probabilities (Chavira & Darwiche, 2008; Dal & Lucas, 2017).

Conjunctive Normal Form (CNF) from logic, where formulas consist of conjunctions of subformulas of literals with only disjunctions, called *clauses*, is commonly used to facilitate compilation. We create for every $X_v \in X_V$ a set of atoms $a(X_v) = \{x_1, \ldots, x_n\}$. Semantically, $x_i \in a(X_v)$ represents X_v being equal its i^{th} value. In addition, an atom ω_j is introduced for every unique probability in X_v's CPT, i.e., ω_j can refer to multiple distinct entries in X_v's CPT if they represent the same probability. A clause is introduced for each entry of the CPT, with an ω_j atom that has a weight $W(\omega_j)$ that is linked to the actual probability, and $W(\overline{\omega_j}) = 1$. Finally clauses are added to prevent inconsistent representations, such as making sure that a variable cannot get multiple values at the same time. This is illustrated by an example (Dal & Lucas, 2017)

Example 2.2 (Bayesian Network encoding)**.** *Let BN $\mathcal{B}$ be defined for variables $\{X, Y\}$ with factorization $P(X, Y) = P(Y \mid X)P(X)$. For simplicity's sake, we focus on just variable X; X has three values, thus the CPT has 3 entries, in this case only two distinct. To encode X and its probabilities we create atoms $a(X) = \{x_1, x_2, x_3\}$. The atom ω_1 is introduced for $X = 1$, ω_2 for $X = 2$ and $X = 3$ (as they have the same probability), with $W(\omega_1) = 0.8$, $W(\omega_2) = 0.1$.*

The CNF representation is as follows:

$$(x_1 \vee x_2 \vee x_3) \wedge (\overline{x_1} \vee \overline{x_2}) \wedge (\overline{x_1} \vee \overline{x_3}) \wedge (\overline{x_2} \vee \overline{x_3}) \wedge (\overline{x_1} \vee \omega_1) \wedge (\overline{x_2} \vee \omega_2) \wedge (\overline{x_3} \vee \omega_2)$$

The first clause enumerates the possible values of X, whereas the second to fourth clause ensure that X (as a random variable) cannot have more than one value. The last three clauses link a probability to an actual value of X. The encoding includes the truth assignments (models) for variable X as shown in Table 1. Note that the weighted model count sums to 1.0 *for this selection of models. However, there are other*

models of this CNF, e.g., model $\{x_1,\overline{x_2},\overline{x_3},\omega_1,\omega_2\}$*, model* $\{\overline{x_1}, x_2, \overline{x_3}, \omega_1, \omega_2\}$*, etc. Only minimal models sum to* 1.0*, i.e., models with the largest number of negations.*

2.2.3. Compilation

Now that we have an encoding, we can consider its compilation to a *Weighted Positive Binary Decision Diagrams* (WPBDD) (Dal & Lucas, 2017). A WPBDD is an ordered BDD that represents a concise factorization of a Boolean formula f as a (rooted) directed acyclic graph with decision nodes, and two terminal nodes labeled with $\top$ (true) and $\bot$ (false). Each non-terminal node v is labeled with a Boolean variable $\text{var}(v) = x_v$ and has two children, $\text{high}(v)$ and $\text{low}(v)$, with a set of weight variables $\text{weights}(v)$ at the edge to node $\text{high}(v)$ (explaining the adjective 'positive' in WPBDD). Each root-terminal path contains a variable at most once, and in a particular total or partial order.

A CNF encoding as described above acts as an entry point for the language compiler (Dudek, Phan, & Vardi, 2020). Such compilers target different variations of DDs.

The respective DD is built using the typical bottom-up strategy (Bryant, 1986), by applying DD operations to construct a DD representing the encoded formula from the previous step. The process of compiling into a respective DD is by far the most expensive operation, compared to the inference step, which is linear in the size of the DD as desired.

2.2.4. Inference

Inference is performed through *Weighted Model Counting* on the DD, WMC for short (Chavira & Darwiche, 2008; Darwiche & Marquis, 2021). This process sums the weight of every truth assignment. In the decision diagram, these assignments are represented by paths and the weights by edge labels. Edges to nodes $\text{high}(v)$ and $\text{low}(v)$ are solid $\rightarrow$ and dashed $\dashrightarrow$, respectively (see below). Since these paths often overlap in the DD structure, inference through model counting is linear in the size of the target representation (Darwiche & Marquis, 2002).

Let's look at a WPBDD compilation and inference example. A WPBDD exactly represents the encoding provided. In order to perform inference we can trivially transform the logical circuit that the WPBDD represents into an arithmetic circuit.

Example 2.3 (Compilation and inference)**.** *Consider again variable* X *from Example 2.2. For the compiled DD the ordering for variable* X *is ordering* $x_3 \prec x_2 \prec x_1$*. Reduction rules specific to WPBDDs allow the removal of the* x_2 *node to further reduce its size. Each path from the root to the* $\top$*-terminal semantically implies evidence. There are three possible paths shown below. If we have evidence prior to traversing the compiled representation, we only consider the paths that are consistent with the evidence.*

Path	*Logic*	*Semantics*
$x_3 \rightarrow \top$	$\overline{x_1} \wedge \overline{x_2} \wedge x_3$	$X = 3$
$x_3 \dashrightarrow x_2 \rightarrow \top$	$\overline{x_1} \wedge x_2 \wedge \overline{x_3}$	$X = 2$
$x_3 \dashrightarrow x_2 \dashrightarrow x_1 \rightarrow \top$	$x_1 \wedge \overline{x_2} \wedge \overline{x_3}$	$X = 1$

To perform inference, we need to link to the probabilities that allows us to compute $P(X = 3) = 0.1$*, by the assignment* $(x_1, x_2, x_3) = (\bot, \bot, \top)$.

The tool PARAGNOSIS offers important ways to optimize compilation and inference by partitioning and parallelization (for details see (Dal et al., 2021, 2023)).

2.3. Mapping to a Bayesian Network

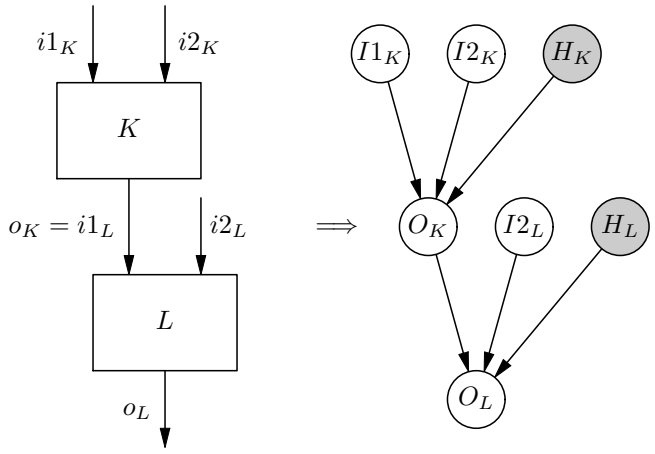
To add a probabilistic aspect to consistency-based diagnosis, a logical diagnostic problem can be mapped to a Bayesian network. There are different ways for translating a diagnostic problem into a *Bayesian diagnostic problem*. Two of these will be highlighted. One of these is the traditional method proposed by Pearl (Pearl, 1988), and implemented later by Srinivas (Srinivas, 1994), and the other is a more recent adaptation introduced by us. The latter one is used for this research, but since it is based on the traditional method both will be expanded upon.

2.3.1. Pearl's Method

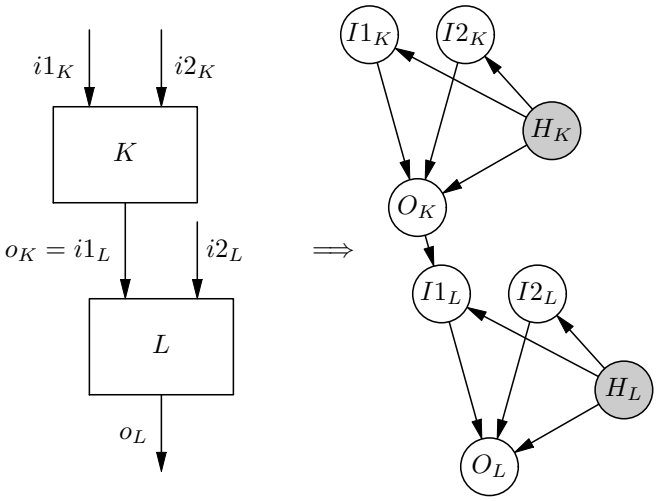
Following Pearl's method (Pearl, 1988), in Fig. 2a an abstract 2-component system has been translated into a Bayesian network. Each input and output of a component are modeled as nodes. A component is modeled as an output node that is the child of all its input nodes. Note that one of the inputs of component L is the output of component K and thus the output node of K is directly linked to the output node of L. Next to these, per component, a health node H (also called 'abnormality node' in the literature) is added as the parent of the output node. This node corresponds to the abnormality predicate in MBD and similarly indicates whether or not the component behaves abnormally: the abnormality literal of the traditional approach could be mimicked by assigning the values 'normal', corresponding logically to the assumption '$\neg\text{Ab}(c)$', whereas the value 'abnormal' would correspond to '$\text{Ab}(c)$'. However, a disadvantage of this approach is that health nodes and input nodes are independent; they only become conditionally dependent when a common child of input nodes or a descendant of the child is instantiated to a value (Pearl, 1988).

2.3.2. Method Based on Connected Health Nodes

The method which is used in the present paper no longer assumes that inputs and health nodes are independent. The health nodes support enforcing extra dependencies between the inputs and outputs, and when none of the children or descendants of the children of the input nodes are instantiated. This method is illustrated in Fig. 2b. Note that in this repre-



(a) Pearl's method with health nodes; health nodes shaded.



(b) New method with health nodes connecting inputs and output; health nodes are shaded.

Figure 2. Two methods of mapping of a simple system model with two components K and L to a Bayesian network.

sentation, all inputs are explicitly represented in the Bayesian network and the relationship between the output of the previous subpart (e.g. component K) and the next subpart (e.g. component L) is represented by an arc between the output node and one input node (e.g. the arc $O_K \rightarrow I1_L$). The conditional probability distribution of any connected output node O, $P(O \mid I_1, \ldots, I_n, H)$, is such that $P(O \mid I_1, \ldots, I_n, H = \text{normal}) \in \{0, 1\}$, dependent on the value of input variables I_k, with $k = 1, \ldots, n$.

2.3.3. Establishing a Bayesian Diagnosis

With the logical diagnostic problem mapped to a Bayesian diagnostic problem, probabilistic inference methods can be used to derive whether or not the components behave correctly as expected to form a diagnosis (Pearl, 1988).

Before such derivation can take place, first the evidence, consisting of observed values of inputs and specific outputs, should be included, which is analogous to the observations 'Obs' in MBD. With probabilistic inference methods, the posterior probabilities of each of the chosen health nodes can be calculated. Then for a given set of health variables H, a *diagnosis* is defined as follows:

$$D = \arg\max_h P(H = h \mid \text{Evidence}, \text{Health-assumptions}) \tag{4}$$

i.e., the assignment h to H with the maximum probability, where it is possible to condition on the health variables not included in H by 'Health-assumptions', the other health variables that are given an (assumed or observed) value. We call this process *assumption-based Bayesian model-based diagnosis*, or Bayesian MBD for short.

The same method can be used for Pearl's Bayesian-network structure of a diagnostic model. However, in that case it is mandatory to instantiate the last output variables to enforce dependence between input and health values and one can no longer simulate various flow schemes.

3. METHODOLOGY

To investigate whether weighted model counting with partitioning offers a suitable and fast algorithm for Bayesian model-based diagnosis, some Bayesian MBD models were designed. Unfortunately, it is virtually impossible to get access to large industrial MBD systems with their associated data for a publication, because of the industry's fear of disclosure of competition-sensitive information to the public domain. For this reason, we had to resort to designing an artificial model, that nevertheless was inspired by existing pipe systems as used in the chemical and oil industry.

The research question that is explored in the remainder of this paper is whether partitioned positive weighted model counting can effectively deal with large Bayesian MBD models in such a way that acceptable diagnostic results are obtained. For this purpose the PARAGNOSIS toolkit was implemented (Dal et al., 2023).

3.1. Basic Elements

Bayesian MBD requires the development of Bayesian-network models of systems, where the models consist of *components*, where some of those components are identical in nature, and compositional ways to merge these models together to build an overall model of a system. The result will be an abstraction of the real-world system that reflects both structure and behavior of the real-world system and that can be used for simulation purposes. In addition for MBD it is necessary to include *behavior modes*, e.g., whether the component is behaving normally or abnormally (other modes are sometimes also used) for each of the components that could be defective, which will be represented as *health variables*. Finally,

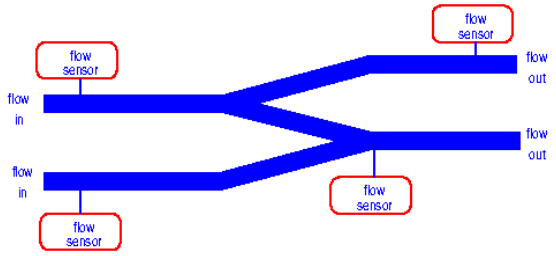


Figure 3. System of connected fluid pipes with one join, one split, and four flow sensors.

information about the behavior of a component is obtained through *sensors*.

3.2. Generating a System

Consider a pipe system, such as the one shown in Fig. 3, that consists of connected pipes, joints of pipes, and splittings of pipes (splits) that transport a fluid, e.g., water or oil, and have an input and output of fluid. The flow of the fluid is measured by means of *flow sensors* and is discretized into three states: 'maxflow', 'lowflow', and 'noflow'. Each pipe has one flow coming into it and one flow going out of it; if pipe x behaves normally:

$$\forall x(\text{PipeIn}(x) = v \rightarrow \text{PipeOut}(x) = v)$$

with v a free variable standing for a state. However, the actual health status of the pipe is ignored here, and assumed to be normal. Alternatively, if we want to take into account the health status and pipe x has a leak, the outgoing flow will be lower than the incoming flow:

$$\forall x((\text{PipeIn}(x) = \text{maxflow} \quad \wedge \quad \text{Health}(x) = \text{leak}) \rightarrow \text{PipeOut}(x) = \text{lowflow}))$$

whereas for normal health we get:

$$\forall x((\text{PipeIn}(x) = v \quad \wedge \quad \text{Health}(x) = \text{normal}) \rightarrow \text{PipeOut}(x) = v))$$

Thus, it is needed to include the health status as an additional condition. Similar logical specifications can be developed for the other elements of pipe systems, i.e., the joints and splittings.

As we use individual pipe components to develop (generate) pipe systems of various complexity and size, as is also done when developing real-life pipe systems, we will number pipe components from input to output flow of the entire system, in terms of two parameters: width (from left to right) and height (top to bottom) of the directed graph (starting with 1). We come back to this issue in Section 4.

A clear disadvantage is that uncertainty in the health status and the likelihood that a leak may give rise to low or no flow is missing in the logical representation. Bayesian model-based diagnosis will support representing this important aspect of diagnosis, i.e., the ability to deal with uncertainty, as will be discussed below.

3.3. Uncertainty and Bayesian-network Components

To model a system that incorporates uncertainty, we need design Bayesian-network components that correspond to the various parts of the real system. Based on the mentioned specifications in the section above, we distinguish pipe, join, and split Bayesian-network components. The health status of a pipe component is controlled by a health variable, called variously 'JoinHealth', 'PipeHealth', 'SplitHealth'. As above, the continuous flow is mapped to discrete values, we distinguish three flow values: maximal (maxflow), low (lowflow), or absent (noflow).

The three basic Bayesian-network components are depicted in Fig. 4, 5, and 6. These can be put together in various topologies giving rise to a plethora of pipe systems. In addition sensors can be added to components to measure the status of the flow in the individual pipes. Sensor readings will be used below as evidence in the experiments to diagnose faults in the different pipe systems. Software and figures have been generated using the R language (R Core Team, 2024), the R-library bnlearn (Scutari, 2024), with GeNIe Modeler[2] for producing graphical figures.

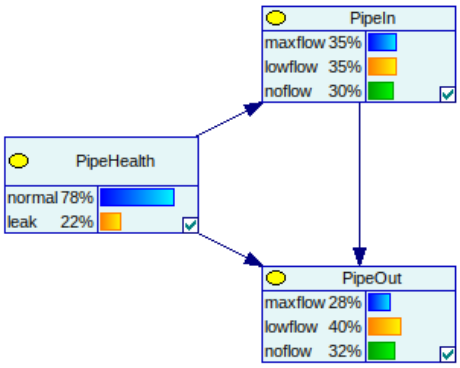


Figure 4. Pipe-shaped component; there is no need to model the actual pipe as it is sufficient to represent the input and output of the pipe and how these relate to each other by means of health modes.

3.4. Compositionality by Probability Distributions

Each pipe is modeled as a flow out of 'PipeOut' with one flow coming into 'PipeIn'. The prior probability of a pipe's health being normal is set to 0.8 ($P(\text{PipeHealth} = \text{normal}) = 0.8$). The distribution of the input of the pipe depends on the output of the previous component in a deterministic manner,

[2] https://www.bayesfusion.com/genie

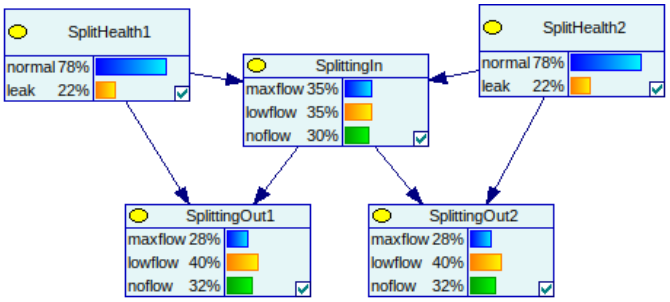


Figure 5. Split-shaped component consisting of a single input and in this case two outputs. In addition, each output is controlled by a health mode.

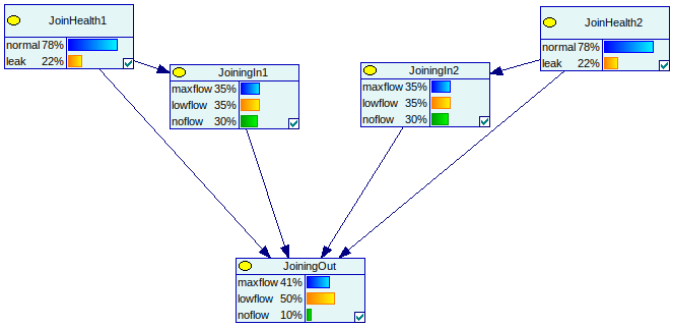


Figure 6. Join-shaped component consisting of two (or more) inputs and a single output that merges the inputs. In addition, the actual merge is controlled by one health node per input.

i.e., we do not consider defects between components. Since this structure is the same for every pipe, the CPT of each pipe is also identical such that the probability distribution

$$P(\text{PipeOut}_i \mid \text{PipeIn}_i = v, \text{PipeHealth}_i = w)$$

is the same for each i, but varies for specific values of v and w as shown in Table 2. In case the component behaves normally, there is no impact on the flow of the pipes, which reflects the logical specifications of Section 3.2. Furthermore, the CPT reflects the impact of a leaking pipe. Given a normal flow, if the pipe is leaking, there is a high probability the flow is reduced, and in severe cases, it may even lead to no flow in the output. If there is low flow in the input, we assume

Table 2. CPT for the output node of a pipe.

		PipeOut$_i$		
		maxflow	**lowflow**	**noflow**
PipeIn$_i$	**PipeHealth$_i$**			
maxflow	**normal**	1	0	0
	leak	0.1	0.8	0.1
lowflow	**normal**	0	1	0
	leak	0	0.8	0.2
noflow	**normal**	0	0	1
	leak	0	0	1

that a leaking pipe has less impact and there is still a high probability there is a low flow in the output, though it may be diminished to no flow.

The split-shaped component is modeled in a similar manner to a pipe, with the exception that this component contains multiple health and output nodes. That is, the prior distributions $P(\text{SplitHealth}_i) = P(\text{PipeHealth}_j)$, the distribution of the 'splittingIn' is determined by the output in its parent component, and

$$P(\text{SplittingOut}_i|\text{SplittingIn} = v, \text{SplitHealth}_i = w) = P(\text{PipeOut}_j|\text{PipeIn}_j = v, \text{PipeHealth}_j = w)$$

for values of i, j (instances of the named components), v and w.

4. Experiments and Evaluation

We have checked the validity of the proposed models in order to prove their usefulness. For the purpose of evaluating the diagnostic capabilities of our weighted model counting method, the 3rd aim of the research, we have generated pipe Bayesian networks 'pipes-sensors-w-h' with height h and width w as described in Section 3.2. This results in w parallel pipes that run h layers deep. Each pipe in the first layer is directly connected to one pipe in the second layer downstream, and so on, for h layers. This effectively lengthens the pipes in the first layer, and creates w parallel pipes of length h. Each pipe in every layer has a sensor that registers *flow* or *noflow*, and is identified by the coordinate in its name. For instance, 'sensor2_3' is the sensor attached to the second pipe in layer 3.

4.1. Experimental Setup

We report some experimental results of our software tool PARAGNOSIS and various pipe Bayesian MBD models. All below experiments ran on a system with as CPU an Intel Hexa Core i7-8750H (2.20-4.10Ghz), 9Mb cache 45W, with Kingston HyperX 16Gb (2× 8Gb) DDR4 2400Mhz RAM.

Evidence is only set on 'FlowOut', 'FlowIn', and (some of the) sensors. Only consistent evidence is considered. This means that 'sensor2_i' $\geq$ 'sensor2_j', for $i < j$. In other terms, it is impossible for an upstream pipe to have no flow and for a downstream pipe to have flow at the same time. We also only consider FlowIn $>$ FlowOut. When we say *maxflow* to *lowflow* or *maxflow* to *noflow*, this means from FlowIn = *maxflow* to FlowOut = *lowflow*, or to FlowOut = *noflow*, respectively.

Consider network 'pipes-sensors-5-2' (5 parallel pipes, 2 layers deep). Table 3 shows diagnostic results for *maxflow* to *noflow*, Table 4 for *maxflow* to *lowflow* and Table 5 for *lowflow* to *noflow*. To allow quick reading we used the following abbreviations: 'pH' stand for 'pipeHealth'; 'jH' for 'joinHealth'; 'sH' for 'splitHealth', respectively. We compute the

Table 3. Results for network 'pipes-sensors-5-2' for maxflow to noflow. Posteriors are computed for health nodes and indicate the probability of a leak. Probabilities preceded by '+' indicate the posterior's increase compared to the prior. The following abbreviations are used: 'pH' stands for 'pipeHealth'; 'jH' for 'joinHealth'; 'sH' for 'splitHealth'.

Sensors set to noflow	**Remaining sensors are set to flow**		**Remaining sensors are not set**	
	3 variables with highest posteriors	3 variables with most increased posteriors	3 variables with highest posteriors	3 variables with most increased posteriors
sensor1_1	pH1_1 (0.738)	pH1_1 (+0.514)	pH1_1 (0.740)	pH1_1 (+0.516)
sensor1_2	jH2 (0.390)	sH1 (+0.201)	sH1 (0.388)	sH1 (+0.198)
	jH3 (0.390)	pH1_2 (+0.024)	jH2 (0.380)	pH1_2 (+0.024)
sensor2_1	pH2_1 (0.738)	pH2_1 (+0.514)	pH2_1 (0.740)	pH2_1 (+0.516)
sensor2_2	jH1 (0.390)	sH2 (+0.201)	sH2 (0.388)	sH2 (+0.198)
	jH3 (0.390)	pH2_2 (+0.024)	jH1 (0.380)	pH2_2 (+0.024)
sensor3_1	pH3_1 (0.738)	pH3_1 (+0.514)	pH3_1 (0.740)	pH3_1 (+0.516)
sensor3_2	jH1 (0.390)	sH3 (+0.201)	sH3 (0.388)	sH3 (+0.198)
	jH2 (0.390)	pH3_2 (+0.024)	jH1 (0.380)	pH3_2 (+0.024)
sensor4_1	pH4_1 (0.738)	pH4_1 (+0.514)	pH4_1 (0.740)	pH4_1 (+0.516)
sensor4_2	jH1 (0.390)	sH4 (+0.201)	sH4 (0.388)	sH4 (+0.198)
	jH2 (0.390)	pH4_2 (+0.024)	jH1 (0.380)	pH4_2 (+0.024)
sensor5_1	pH5_1 (0.738)	pH5_1 (+0.514)	pH5_1 (0.740)	pH5_1 (+0.516)
sensor5_2	jH1 (0.390)	sH5 (+0.201)	sH5 (0.388)	sH5 (+0.198)
	jH2 (0.390)	pH5_2 (+0.024)	jH1 (0.380)	pH5_2 (+0.024)
sensor1_2	pH1_2 (0.579)	pH1_2 (+0.354)	pH1_1 (0.444)	pH1_1 (+0.219)
	jH2 (0.390)	sH1 (+0.039)	pH1_2 (0.444)	pH1_2 (+0.219)
	jH3 (0.390)	jH2 (+0.011)	jH2 (0.380)	sH1 (+0.103)
sensor2_2	pH2_2 (0.579)	pH2_2 (+0.354)	pH2_1 (0.444)	pH2_1 (+0.219)
	jH1 (0.390)	sH2 (+0.039)	pH2_2 (0.444)	pH2_2 (+0.219)
	jH3 (0.390)	jH1 (+0.011)	jH1 (0.380)	sH2 (+0.103)
sensor3_2	pH3_2 (0.579)	pH3_2 (+0.354)	pH3_1 (0.444)	pH3_1 (+0.219)
	jH1 (0.390)	sH3 (+0.039)	pH3_2 (0.444)	pH3_2 (+0.219)
	jH2 (0.390)	jH1 (+0.011)	jH1 (0.380)	sH3 (+0.103)
sensor4_2	pH4_2 (0.579)	pH4_2 (+0.354)	pH4_1 (0.444)	pH4_1 (+0.219)
	jH1 (0.390)	sH4 (+0.039)	pH4_2 (0.444)	pH4_2 (+0.219)
	jH2 (0.390)	jH1 (+0.011)	jH1 (0.380)	sH4 (+0.103)
sensor5_2	pH5_2 (0.579)	pH5_2 (+0.354)	pH5_1 (0.444)	pH5_1 (+0.219)
	jH1 (0.390)	sH5 (+0.039)	pH5_2 (0.444)	pH5_2 (+0.219)
	jH2 (0.390)	jH1 (+0.011)	jH1 (0.380)	sH5 (+0.103)

posteriors of value *leak* for all health variables, given the evidence that a set of sensors is set to *noflow*. This set can be found in the leftmost column. 'FlowOut' and 'FlowIn' are also observed respectively.

We compare two experiments. Columns 2-3 represent the experiment where all sensors are observed. Unobserved sensors (those not present in column 1) are set to *flow*. Columns 4-5 represent the experiment where unobserved sensors are not set. Columns 2 and 4 contain the health variables with the highest *leak* probability, whereas column 3 and 5 contain the health variables with the most increased *leak* posteriors, compared to the posteriors computed with only 'FlowOut' and 'FlowIn' in the evidence.

4.2. Observations and Diagnostic Results

The posteriors in Table 3 (*maxflow* to *noflow*) show that diagnoses are consistent with the evidence. Consider evidence 'sensor1_1' and 'sensor1_2' equal to *noflow* and all remaining sensors are set to *flow*. The most likely location for a leak is 'pipeHealth1_1'. When the evidence does not include values for sensor1_1, we see that pipeHealth1_2 indicates a leak. Resulting diagnoses, consisting of the highest probabilities for the health variables, seemed to correspond to what we expected.

When removing the *flow* sensors from the evidence we see that the evidence 'sensor1_2' does not lead to a definitive leak (a probability greater than 0.5) as indicated by pipeHealth1_2. However, pipeHealth1_2 has the highest leak probability along with upstream pipeHealth1_1. Their posteriors also have increased the most as indicated in the last column. This finding is also logical, we have not set 'sensor1_1' to *flow*, thus the leak can still be in any layer.

For Table 4 and Table 5 we see the same behavior, thereby validating the diagnostic capabilities of our Bayesian MBD approach.

4.3. Larger Networks

We have created larger systems using the description in Section 4.1, and perform weighted model counting using PARAG-

Table 4. Results for network 'pipes-sensors-5-2' for maxflow to lowflow. Posteriors are computed for health nodes and indicate the probability of a leak. Probabilities preceded by '+' indicate the posterior's increase compared to the prior. The following abbreviations are used: 'pH' stands for 'pipeHealth'; 'jH' for 'joinHealth'; 'sH' for 'splitHealth'.

Sensors set to noflow	**Remaining sensors are set to flow**		**Remaining sensors are not set**	
	3 variables with highest posteriors	3 variables with most increased posteriors	3 variables with highest posteriors	3 variables with most increased posteriors
sensor1_1	pH1_1 (0.752)	pH1_1 (+0.522)	pH1_1 (0.750)	pH1_1 (+0.520)
sensor1_2	sH1 (0.388)	sH1 (+0.201)	sH1 (0.386)	sH1 (+0.199)
	jH2 (0.264)	pH1_2 (+0.021)	pH1_2 (0.251)	pH1_2 (+0.020)
sensor2_1	pH2_1 (0.752)	pH2_1 (+0.522)	pH2_1 (0.750)	pH2_1 (+0.520)
sensor2_2	sH2 (0.388)	sH2 (+0.201)	sH2 (0.386)	sH2 (+0.199)
	jH1 (0.264)	pH2_2 (+0.021)	pH2_2 (0.251)	pH2_2 (+0.020)
sensor3_1	pH3_1 (0.752)	pH3_1 (+0.522)	pH3_1 (0.750)	pH3_1 (+0.520)
sensor3_2	sH3 (0.388)	sH3 (+0.201)	sH3 (0.386)	sH3 (+0.199)
	jH1 (0.264)	pH3_2 (+0.021)	pH3_2 (0.251)	pH3_2 (+0.020)
sensor4_1	pH4_1 (0.752)	pH4_1 (+0.522)	pH4_1 (0.750)	pH4_1 (+0.520)
sensor4_2	sH4 (0.388)	sH4 (+0.201)	sH4 (0.386)	sH4 (+0.199)
	jH1 (0.264)	pH4_2 (+0.021)	pH4_2 (0.251)	pH4_2 (+0.020)
sensor5_1	pH5_1 (0.752)	pH5_1 (+0.522)	pH5_1 (0.750)	pH5_1 (+0.520)
sensor5_2	sH5 (0.388)	sH5 (+0.201)	sH5 (0.386)	sH5 (+0.199)
	jH1 (0.264)	pH5_2 (+0.021)	pH5_2 (0.251)	pH5_2 (+0.020)
sensor1_2	pH1_2 (0.653)	pH1_2 (+0.423)	pH1_1 (0.459)	pH1_1 (+0.228)
	jH2 (0.283)	sH1 (+0.028)	pH1_2 (0.459)	pH1_2 (+0.228)
	jH3 (0.283)	jH2 (+0.015)	sH1 (0.292)	sH1 (+0.105)
sensor2_2	pH2_2 (0.653)	pH2_2 (+0.423)	pH2_1 (0.459)	pH2_1 (+0.228)
	jH1 (0.283)	sH2 (+0.028)	pH2_2 (0.459)	pH2_2 (+0.228)
	jH3 (0.283)	jH1 (+0.015)	sH2 (0.292)	sH2 (+0.105)
sensor3_2	pH3_2 (0.653)	pH3_2 (+0.423)	pH3_1 (0.459)	pH3_1 (+0.228)
	jH1 (0.283)	sH3 (+0.028)	pH3_2 (0.459)	pH3_2 (+0.228)
	jH2 (0.283)	jH1 (+0.015)	sH3 (0.292)	sH3 (+0.105)
sensor4_2	pH4_2 (0.653)	pH4_2 (+0.423)	pH4_1 (0.459)	pH4_1 (+0.228)
	jH1 (0.283)	sH4 (+0.028)	pH4_2 (0.459)	pH4_2 (+0.228)
	jH2 (0.283)	jH1 (+0.015)	sH4 (0.292)	sH4 (+0.105)
sensor5_2	pH5_2 (0.653)	pH5_2 (+0.423)	pH5_1 (0.459)	pH5_1 (+0.228)
	jH1 (0.283)	sH5 (+0.028)	pH5_2 (0.459)	pH5_2 (+0.228)
	jH2 (0.283)	jH1 (+0.015)	sH5 (0.292)	sH5 (+0.105)

NOSIS (Dal et al., 2023). It is clear that the created systems are particularly strenuous on the inference side, due to the joining node in the network. Its CPT increases exponentially when width w increases. Table 6 shows compilation and inference times of these networks. The results reflect the aforementioned comment, as compilation and inference time most notably increase as the width of the network increases. As a comparison, the well known Munin network (Jensen & Andreassen, 2008) is considered to be large, and has 1041 variables with 98423 probabilities (Dal et al., 2021). This demonstrates the inference capabilities of weighted model counting using PARAGNOSIS (Dal et al., 2023). Fig. 7 show a nearly linear increase in compilation time with respect to the number of probabilities in the networks. However, the increase in compilation and inference time is more exponential in nature as we increase the width of the network.

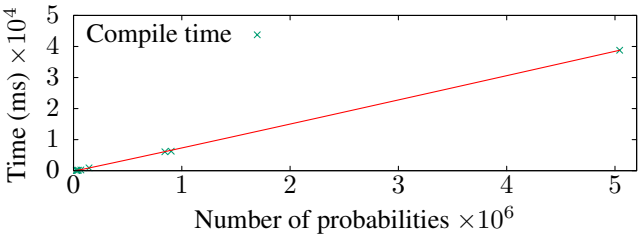


Figure 7. Compilations time with respect to the number of probabilities.

5. DISCUSSION AND CONCLUSIONS

Following the analysis of the diagnostic behavior, it appears that the compositional Bayesian-network structures developed in the sections above display behavior that is at least partly natural and intuitive. For example, a Bayesian model-based diagnostic model of a pipe system that has no input flow will raise an inconsistency in its (conditional) probability distribution if its output flow is assumed to be low or maximal. Another interesting aspect of the behavior is that faults are assumed to be closest (in terms of path length) to the entered observations that are incompatible with the expected behavior. This is a consequence of the propagation of probabilistic

Table 5. Results for network 'pipes-sensors-5-2' for lowflow to noflow. Posteriors are computed for health nodes and indicate the probability of a leaku. Probabilities preceded by '+' indicate the posterior's increase compared to the prior. The following abbreviations are used: 'pH' stands for 'pipeHealth'; 'jH' for 'joinHealth'; 'sH' for 'splitHealth'.

Sensors set to noflow	**Remaining sensors are set to flow**		**Remaining sensors are not set**	
	3 variables with highest posteriors	3 variables with most increased posteriors	3 variables with highest posteriors	3 variables with most increased posteriors
sensor1_1 sensor1_2	sH1 (0.506) pH1_1 (0.498) jH2 (0.360)	pH1_1 (+0.293) sH1 (+0.216) pH1_2 (+0.042)	pH1_1 (0.508) sH1 (0.500) jH2 (0.356)	pH1_1 (+0.303) sH1 (+0.210) pH1_2 (+0.044)
sensor2_1 sensor2_2	sH2 (0.506) pH2_1 (0.498) jH1 (0.360)	pH2_1 (+0.293) sH2 (+0.216) pH2_2 (+0.042)	pH2_1 (0.508) sH2 (0.500) jH1 (0.356)	pH2_1 (+0.303) sH2 (+0.210) pH2_2 (+0.044)
sensor3_1 sensor3_2	sH3 (0.506) pH3_1 (0.498) jH1 (0.360)	pH3_1 (+0.293) sH3 (+0.216) pH3_2 (+0.042)	pH3_1 (0.508) sH3 (0.500) jH1 (0.356)	pH3_1 (+0.303) sH3 (+0.210) pH3_2 (+0.044)
sensor4_1 sensor4_2	sH4 (0.506) pH4_1 (0.498) jH1 (0.360)	pH4_1 (+0.293) sH4 (+0.216) pH4_2 (+0.042)	pH4_1 (0.508) sH4 (0.500) jH1 (0.356)	pH4_1 (+0.303) sH4 (+0.210) pH4_2 (+0.044)
sensor5_1 sensor5_2	sH5 (0.506) pH5_1 (0.498) jH1 (0.360)	pH5_1 (+0.293) sH5 (+0.216) pH5_2 (+0.042)	pH5_1 (0.508) sH5 (0.500) jH1 (0.356)	pH5_1 (+0.303) sH5 (+0.210) pH5_2 (+0.044)
sensor1_2	pH1_2 (0.376) jH2 (0.360) jH3 (0.360)	pH1_2 (+0.171) jH2 (+0.001) jH3 (+0.001)	sH1 (0.372) jH2 (0.357) jH3 (0.357)	pH1_1 (+0.117) pH1_2 (+0.117) sH1 (+0.081)
sensor2_2	pH2_2 (0.376) jH1 (0.360) jH3 (0.360)	pH2_2 (+0.171) jH1 (+0.001) jH3 (+0.001)	sH2 (0.372) jH1 (0.357) jH3 (0.357)	pH2_1 (+0.117) pH2_2 (+0.117) sH2 (+0.081)
sensor3_2	pH3_2 (0.376) jH1 (0.360) jH2 (0.360)	pH3_2 (+0.171) jH1 (+0.001) jH2 (+0.001)	sH3 (0.372) jH1 (0.357) jH2 (0.357)	pH3_1 (+0.117) pH3_2 (+0.117) sH3 (+0.081)
sensor4_2	pH4_2 (0.376) jH1 (0.360) jH2 (0.360)	pH4_2 (+0.171) jH1 (+0.001) jH2 (+0.001)	sH4 (0.372) jH1 (0.357) jH2 (0.357)	pH4_1 (+0.117) pH4_2 (+0.117) sH4 (+0.081)
sensor5_2	pH5_2 (0.376) jH1 (0.360) jH2 (0.360)	pH5_2 (+0.171) jH1 (+0.001) jH2 (+0.001)	sH5 (0.372) jH1 (0.357) jH2 (0.357)	pH5_1 (+0.117) pH5_2 (+0.117) sH5 (+0.081)

information through the network which reflects the observations combined with the associated uncertainty derived from the conditional probability tables. At first thought, one may think that this behavior is not compatible with the actual real world, as one would expect that for identical components, the likelihood of failure would also be identical independent of where the component is located in the overall system. However, as *only* by observing flow input and output and sensor data allows one to locate faults, the probabilistic reasoning provided by a Bayesian model-based diagnostic model rightfully exploits that information to the maximal extent and does indeed offer information where to look first.

In principle, the Bayesian-network network structure and its associated probabilistic parameters can be learned from observational data of a working real machine, although this will be associated with some major challenges. Where it would be straightforward to learn the prior distribution of health variables from the data, for example for 'PipeHealth', it would be harder to learn the conditional distributions

$$P(\mathrm{PipeOut}|\mathrm{PipeIn} = v, \mathrm{PipeHealth} = w),$$

for values of v, w, from data, partly because these probabilities are supposed to represent generic local probabilities, whereas in real-world systems quite a lot of the local behavior is determined by non-local behavior arising elsewhere in a system. Measurements that fully explain local behavior of components will usually not be available. Instead, available sensor data can be exploited, and basically this requires the development of new methods to answer questions of how local probability distributions can be approximated with the data from real-world systems that *can be* measured. Since parameter learning does not appear straightforward, it can be expected that structure learning will be even harder. However, here one has to keep in mind that for cyberphysical systems there often is quite some knowledge available already about its architecture and functional components, which clearly makes this problem much less challenging. Thus the positive message is that with relatively little effort a good starting point for developing a Bayesian model-based diagnostic model is within reach.

It should be mentioned that the example pipe model which

Table 6. Compilation and inference times for pipes-sensors-w-h, where compilation size is the number of arithmetic operators in the compiled representation.

Network		Number of	Number of	Compilation	Compilation	Maginalization
Width w	Height h	variables	probabilities	size	time (ms)	time (ms)
5	10	223	25833	76908	98.165	4.334
5	20	423	28033	128619	116.489	8.927
5	30	623	30233	107223	112.460	6.582
5	40	823	32433	136836	116.275	9.780
5	50	1023	34633	133992	115.188	7.974
5	100	2023	45633	202545	153.610	11.971
5	200	4023	67633	341820	235.962	20.261
6	10	267	143049	127431	905.657	9.245
7	10	311	843561	498813	6085.595	37.000
7	200	5631	902081	645177	6275.068	42.671
8	10	355	5043465	463594	38781.416	36.785

was employed in the research, has some limitations. In the first place, because of the restriction in our research to probabilistic inference in discrete Bayesian networks. It would have been more natural to use hybrid Bayesian networks to represent the behavior of a pipe system, with continuous variables for the representation of flow and sensor information and discrete for the health variables. Nevertheless, discrete variables do allow one to approximate continuous variables usually to a sufficient degree. Furthermore, in the context of flow modeling, the assumption that flow can be modeled by a directed acyclic graph may be questioned, although propagation of probabilistic information goes in both directions, in and against the direction of the arcs.

The issues mentioned above do not interfere with other conclusions concerning the probabilistic diagnostic method designed and whether exact probabilistic inference is feasible for large diagnostic problems. We have also tested large versions of our model-based diagnostic models, in order to investigate the limits of weighted model counting. We are able to perform weighted model counting in networks that are significantly larger than the largest networks in the field of Bayesian networks (Dal et al., 2023).

All in all, this research contributes to ideas on how Bayesian model-based diagnosis can be applied to large cyberphysical systems. Future research should shed light on whether the proposed probabilistic diagnostic methods can be of value in diagnosing faults in other systems than the pipe systems we studied.

REFERENCES

Bryant, R. E. (1986). Graph-based algorithms for Boolean function manipulation. *Transactions on Computers*, *100*, 677–691.

Chavira, M., & Darwiche, A. (2008). On probabilistic inference by weighted model counting. *Artificial Intelligence*, *172*, 772–799.

Dal, G. H., Laarman, A., & Lucas, P. J. F. (2023). ParaGnosis: A tool for parallel knowledge compilation. In *Model Checking Software: 29th International Symposium, SPIN 2023, Paris, France, April 26–27, 2023, Proceedings* (pp. 22–37).

Dal, G. H., Laarman, A. W., Hommersom, A., & Lucas, P. J. F. (2021). A compositional approach to probabilistic knowledge compilation. *International Journal of Approximate Reasoning*, *138*, 38–66.

Dal, G. H., & Lucas, P. J. F. (2017). Weighted positive binary decision diagrams for exact probabilistic inference. *International Journal of Approximate Reasoning*, *90*, 411–432.

Dal, G. H., Michels, S., & Lucas, P. J. F. (2017). Reducing the cost of probabilistic knowledge compilation. In *Proceedings of Machine Learning Research* (Vol. 73, pp. 141–152).

Darwiche, A., & Marquis, P. (2002). A knowledge compilation map. *Journal of Artificial Intelligence Research*, *17*, 229–264.

Darwiche, A., & Marquis, P. (2021). On quantifying literals in Boolean logic and its applications to explainable AI. *Journal of Artificial Intelligence Research*, *72*, 285–328.

de Kleer, J. (1991). Focusing on probable diagnoses. In *Proc. of AAAI-1991* (pp. 842–848).

de Kleer, J., Mackworth, A. K., & Reiter, R. (1992). Characterizing diagnoses and systems. *Artificial Intelligence*, *56*(2-3), 197–222. doi: 10.1016/0004-3702(92)90027-u

de Kleer, J., & Williams, B. C. (1987). Diagnosing multiple faults. *Artificial Intelligence*, *32*(1), 97–130. doi: 10.1016/0004-3702(87)90063-4

Dowdeswell, B., Sinha, R., & MacDonell, S. G. (2020). Finding faults: A scoping study of fault diagnostics for Industrial Cyber–Physical Systems. *Journal of Systems and Software*, *168*, 110638. doi:

10.1016/j.jss.2020.110638

Dudek, J. M., Phan, V., & Vardi, M. Y. (2020). ADDMC: Weighted model counting with algebraic decision diagrams. In *International Conference on Artificial Intelligence* (pp. 1468–1476).

Genesereth, M., & Nilsson, N. (1987). *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann.

Grievink, G. (2022). *Model-based probabilistic diagnostics of cyber-physical systems.* Bachelor Thesis, University of Twente, The Netherlands.

Jensen, K., & Andreassen, S. (2008). Generic causal probabilistic networks: A solution to a problem of transferability in medical decision support. *Computer Methods and Programs in Biomedicine*, *89*(2), 189–201. doi: 10.1016/j.cmpb.2007.10.015

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lee, E. A. (2008). Cyber physical systems: Design challenges. In *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)* (pp. 363–369). doi: 10.1109/ISORC.2008.25

Lucas, P. J. F. (1996). *Structures in Diagnosis: from theory to clinical application* (Published PhD Thesis). Free University (VU) of Amsterdam, The Netherlands.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan kaufmann.

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org`

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, *32*(1), 57–95. doi: 10.1016/0004-3702(87)90062-2

Ruijters, E., & Stoelinga, M. (2015). Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. *Computer Science Review*, *15-16*, 29–62. doi: 10.1016/j.cosrev.2015.03.001

Scutari, M. (2024). bnlearn – an R package for Bayesian network learning and inference [Computer software manual]. Retrieved from `https://bnlearn.com`

Srinivas, S. (1994). A probabilistic approach to hierarchical model-based diagnosis. In *Uncertainty in Artifical Intelligence* (pp. 538–545). Elsevier.

MOXAI – Manufacturing Optimization through Model-Agnostic Explainable AI and Data-Driven Process Tuning

Clemens Heistracher[1]*, Anahid Wachsenegger[2]*, Axel Weißenfeld[2], and Pedro Casas[2]

[1] *craftworks GmbH, Vienna, Austria*
clemens.heistracher@craftworks.at

[2] *AIT Austrian Institute of Technology, Vienna, Austria*
name.surname@ait.ac.at

ABSTRACT

Modern manufacturing equipment offers numerous configurable parameters for optimization, yet operators often underutilize them. Recent advancements in machine learning (ML) have introduced data-driven models in industrial settings, integrating key equipment characteristics. This paper evaluates the performance of ML models in classification tasks, revealing nuanced observations. Understanding model decision-making processes in failure detection is crucial, and a guided approach aids in comprehending model failures, although human verification is essential. We introduce *MOXAI*, a data-driven approach leveraging existing pre-trained ML models to optimize manufacturing machine parameters. MOXAI underscores the significance of explainable artificial intelligence (XAI) in enhancing data-driven process tuning for production optimization and predictive maintenance. MOXAI assists operators in adjusting process settings to mitigate machine failures and production quality degradation, relying on techniques like DiCE for automatic counterfactual generation and LIME to enhance the interpretability of the ML model's decision-making process. Leveraging these two techniques, our research highlights the significance of explaining the model and proposing the recommended parameter setting for improving the process.

1. INTRODUCTION

Today's highly automated manufacturing equipment often provides many configurable parameters to ensure optimal production and accommodate an increased range of products. In practice, machine operators and process engineers rely on a limited set of well-understood key parameters for process controlling and optimization, overlooking the broader space of configurable options and underutilizing the potential to enhance equipment effectiveness. The increased demand for individualization and, consequently, the decrease in batch sizes amplify this effect and further increase the workload for operators. Recent advances in machine learning have led to a surge in data-driven AI/ML models deployed in industrial scenarios for applications such as quality inspection and predictive maintenance, which have integrated key characteristics and patterns of production equipment.

The demand for explainability becomes crucial to optimizing complex manufacturing and production processes as models grow more intricate, resembling "black boxes" that hinder users from understanding the rationale behind predictions. Explainable Artificial Intelligence (XAI) methods address this challenge by providing human-understandable explanations for data-driven decisions. In XAI, two primary categories are evident (Molnar, 2020): model-agnostic and model-specific approaches. Model-agnostic techniques, such as feature importance and surrogate models, offer insights into decision-making processes across various models. Conversely, model-specific methods delve into a model's intrinsic aspects, such as coefficients in linear regression or visualizing decision cuts in decision trees. Local and global scopes characterize explanations, with techniques like Local Interpretable Model-Agnostic (LIME) (Ribeiro, Singh, & Guestrin, 2016), and Shapely Additive Explanations (Lundberg & Lee, 2017) offering local insights. Another popular approach in XAI is counterfactual explanations (Ates, Aksar, Leung, & Coskun, 2021; Jalali, Haslhofer, Kriglstein, & Rauber, 2023), which determine changes to input data necessary for altering a model's output.

In this work, we strive to automate the process of providing recommendations to machine operators in an interpretable manner, empowering them to understand and adjust parameters effectively for optimal performance (Fig. 1). For this purpose, we introduce *MOXAI* tuning, a data-driven approach

leveraging existing pre-trained AI/ML data models to optimize manufacturing machine parameters by applying model-agnostic counterfactual explanations. Given that numerous manufacturing optimization and predictive maintenance tasks are framed within a binary classification –distinguishing between healthy and damaged assets, or regression – predicting health indicators or remaining useful life, counterfactuals emerge as a compelling solution.

To demonstrate the concept and applicability of the proposed approach, we apply MOXAI on the AI4I 2020 Predictive Maintenance Dataset (Matzka, 2020), which is a simulated dataset designed to mirror authentic predictive maintenance data typically observed in industrial manufacturing settings. Applying MOXAI to those samples where the model predicts machine failures, we can analyze the rationale behind these predictions and obtain suggested modifications to fine-tune different process parameters and prevent machine failures. By querying MOXAI for explanations, we assume that mitigating the reasons behind the model failure prediction will result in an enhanced quality outcome. MOXAI explanations are constrained to a subset of features directly or indirectly controlled by the operator. Evaluations involve comparing model suggestions with production settings to quantify the impact on machine failures, by applying LIME to verify the explanations discovered.

The rest of the paper is structured as follows: First, an overview of the related work is given in Section 2 . Section 3 introduces MOXAI. Section 4 describes the experimental setup considered for evaluation purposes, presenting experimental results. Further discussion on results and MOXAI's approach is presented in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

Two main categories emerge in the domain of XAI: model-agnostic and model-specific approaches. Model-agnostic techniques do not rely on specific model characteristics and can generally be applicable across various models to provide insights into their decision-making process (Molnar, 2020). Such methods include feature importance (e.g., shapely values (Lipovetsky & Conklin, 2001)) and model approximation techniques (e.g., surrogate models (Ribeiro et al., 2016)). Conversely, model-specific approaches study the intrinsic aspects of a model and offer a deeper understanding of its learning structure. For instance, coefficients of a linear regression model, visualization of decision cuts of a shallow decision tree, or more complex approaches such as Layer-Wise propagation explanations (Bach et al., 2015), DeepLift (Shrikumar, Greenside, & Kundaje, 2017), and Class Activation Map (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016), which visualize the distributed weights of a neural network.

The scope of the explanations provided by either of the aforementioned techniques can be either local (explaining one sample) or global (explaining all the samples) (Molnar, 2020). Local Interpretable Model-Agnostic (LIME) (Ribeiro et al., 2016), and Shapely Additive exPlanations (Lundberg & Lee, 2017) are popular techniques that produce local explanations. Recent studies suggest that the comprehensibility of local explanations, specifically when including the counterfactuals, increases the human understanding of the model's decision boundary (Jalali et al., 2023). Counterfactual explanations are "hypothetical samples that are as similar as possible to the sample that is explained while having a different classification label" (Ates et al., 2021). Therefore, we argue that combining a local explainability approach with generating counterfactuals can help an end user understand the small meaningful changes that cause the shift in the model's decision with minimal computational effort.

Many XAI approaches have been applied in the literature to address manufacturing optimization problems. Schockaert et al. (Schockaert, Macher, & Schmitz, 2020) propose an approach for local interpretability of a model optimized on training data, which forecasts the temperature of the hot metal a blast furnace produces. Combining a Variational AutoEncoder (VAE) with LIME significantly improves generated synthetic samples for training the ML model. Seiffer et al. (Seiffer, Ziekow, Schreier, & Gerling, 2021) develop a framework to detect temporal changes in manufacturing data with SHAP values to enhance error prediction. The framework detects and handles concept drift so that the generated ML models are of sufficient quality in the long term. Jakubowski et al. (Jakubowski, Stanisz, Bobek, & Nalepa, 2021) developed an LSTM autoencoder model for detecting anomalies in the hot rolling process to produce steel coils. They applied SHAP explanations to determine the reasons for anomalies. Regarding model interpretability, Jakubowski et al. (Jakubowski, Stanisz, Bobek, & Nalepa, 2022) employed the SHAP method and counterfactual explanations to gain insight into the decisions made by their trained models. These explanations effectively highlighted the features responsible for the abnormal state of the mill or work rolls, helping identify the anomaly's root cause. Ameli et al. (Ameli et al., 2022) employ XAI methodologies to determine the specific sensors exhibiting anomalies, enhancing decision-making within glass production monitoring. These sensors are localized, analyzing the cause of anomalies by saliency XAI. The approach of Senoner et al. (Senoner, Netland, & Feuerriegel, 2022) involves the development of a data-driven decision model by leveraging high-dimensional data with nonlinear relationships alongside SHAP to discern the intricate relationships between production parameters and manufacturing process quality.

In summary, XAI methods are sporadically utilized in production and predictive maintenance to optimize models and

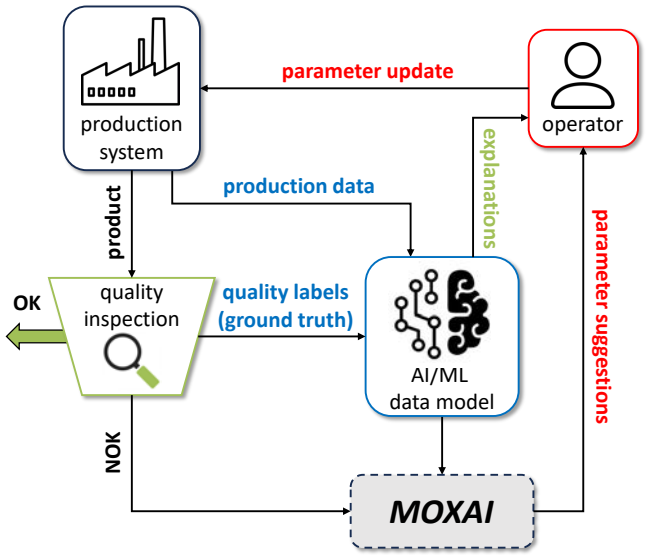


Figure 1. MOXAI information flow for parameter recommendations.

enhance understanding. The approach involving counterfactuals has been minimally employed thus far despite its considerable promise in this domain.

3. Methodology

The general idea of MOXAI is to extract suggestions for *production parameters* (i.e., part of the AI/ML model input features) leading to a desired manufacturing target, relying on pre-trained AI/ML models, and using XAI local explanations through so-called counterfactual instances. A counterfactual instance is a synthetic data point similar to the original instance but with a different model outcome. It is created by perturbing the model's input parameters within certain bounds. Our goal through counterfactual explanations is to answer the question *"what changes to the input (production) parameters of the model would have resulted in a different prediction?"*. MOXAI allows for faster fine-tuning of the configuration of a production process by iterating over instances for which the production output is not as desired.

3.1. Automatic Parameter-Settings' Recommendation

We developed MOXAI to guide machine operators toward better production parameters in case of product quality deviations. We envision a scenario in which process control, quality inspections, or data-driven models indicate a deviation, and the operator is uncertain how to modify the configuration. MOXAI leverages methods from explainable AI to suggest an optimized configuration based on the most recent sample. It requires an existing data model for product quality prediction based on the process parameter and configuration. Our model-agnostic method requires the changeable machine configuration to be part of the model input.

We use the framework for Diverse Counterfactual Explanations (DiCE) (Mothilal, Sharma, & Tan, 2020) to generate recommendations. We leverage counterfactuals to suggest machine parameters that produce good product quality according to the data model. DiCE aims to generate actionable counterfactual sets, ensuring that individual counterfactual examples are feasible and diverse. To achieve this, DiCE adapts diversity metrics through diversity via Determinantal Point Processes (Kulesza, Taskar, et al., 2012) and incorporates feasibility using proximity constraints and user-defined constraints. Process parameters are optimized by extracting the model's capability to determine which parameters lead to a high-quality product. It also addresses sparsity by considering the minimal number of features that must be changed to transition to the counterfactual class. Additionally, it allows users to specify constraints on feature manipulation, such as box constraints on feasible feature ranges, to ensure the practicality of counterfactual examples within real-world constraints. The MOXAI workflow is depicted in Figure 1.

3.2. Human-Guided Correction of Model Failures

To understand the model's decision boundary for detecting defected cases from no-defect cases, we apply LIME, which also offers understandable visualizations for operators and developers to understand why the model failed to detect defective samples. LIME produces instance-based explanations by estimating the decision boundary of the black-box model within a narrow neighborhood. The underlying assumption is that a linear model can effectively approximate the local decision boundary of the black box. The coefficients of this linear model then elucidate the contribution of each feature to the prediction of a sample within this neighborhood. Consequently, LIME's explanations are represented by feature value boundaries. These boundaries signify the impact of each feature; when the feature values fall within these boundaries in a given local neighborhood, they influence the model's decision toward or away from a particular class.

MOXAI's correction algorithm utilizes LIME and examines the top five common explanations provided by this approach for all instances. It performs the following steps: it counts the frequency of these explanations; it then ranks the explanations based on their frequency counts. Next, it records the lowest and highest bounds observed for the most influential feature in the explanations. Human input may be needed from a domain expert who has viewed the data and understands the feature boundaries at this stage. This is necessary because LIME sometimes presents an upper or lower-bound inequality. In such cases, we need to determine the missing boundary. The algorithm then iterates over the generated list and replaces the corresponding feature in the explanations with a randomly generated float within the boundary range. We continue the iteration if this alteration does not rectify the model's prediction. If the alteration corrects the prediction,

Table 1. Results of trained models on the test set.

	Recall	Precision	F1-Score
Nearest Neighbor (KNN)	0.97	0.97	0.97
Decision Tree (DT)	0.99	0.99	0.99
Random Forest (RF)	0.99	0.99	0.99
Gradient Boosting (GBM)	0.99	0.99	0.99
Neural Network (MLP)	0.97	0.94	0.96

Table 2. Detailed results of the trained decision tree on the test set.

	Accuracy	Recall	Precision	F1-score
HDF	0.99	0.95	0.93	0.94
PWF	0.99	0.92	0.89	0.90
OSF	0.99	0.73	0.87	0.78
Machine-Failure	0.99	0.99	0.99	0.99

we proceed to the next misclassified sample. We further emphasize that this approach merely identifies the approximate decision boundary of the model rather than identifying the actual cause of the defect. We can only observe the parameter responsible for the model's misclassification, which may or may not directly correlate with the underlying cause of the defect. The outcomes of this algorithm are discussed in Section 4.3.

4. Experimental Setup and Results

We demonstrate MOXAI's operation using the AI4I dataset, a synthetic dataset commonly used in the scientific community. The AI4I dataset covers a realistic industrial use case and provides an analytical definition for most error types, which can be used to validate corrections as suggested by MOXAI. The dataset consists of 10,000 samples with five numerical features of a milling process, a categorical feature for different product types, and the target variables, which describe the state of five error types:

- Tool wear failure (TWF): The tool fails after a random up-time between 200 - 240 minutes.
- Heat dissipation failure (HDF): The tool fails due to small temperature differences between the tool and air and slow rotational speeds.
- Power failure (PWF): The tool fails for very high or very low power, defined as the product of torque and rotational speed.
- Overstrain failure (OSF): Product variant-dependent error for high tool wear and torque combination.
- Random failures (RNF): A randomly assigned error type.

We exclude TWF and RNF failures from the evaluations due to their random component, as we require an analytical definition of the error for validation.

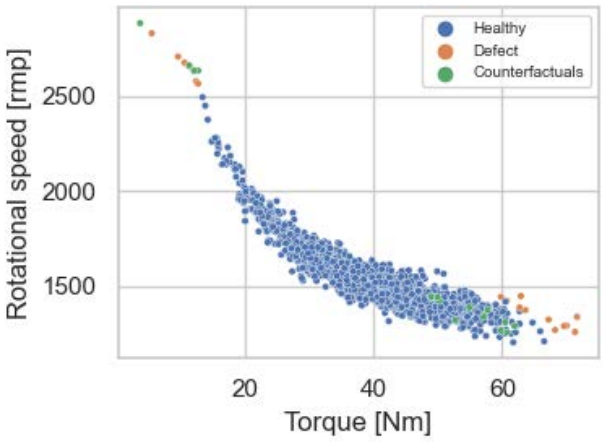


Figure 2. Power Failure (PWF) healthy and defect samples in the test set, as well as generated counterfactual samples, for rotation speed vs. torque of the milling process.

4.1. Data Preprocessing and Modeling

The machine learning model is the core of our approach, and we trained different models following standard best practices. We use a stratified split of 80% of the data for training/validation and 20% for testing, resulting in 7714 samples for the healthy state and 234 defects, of which 115 are HDF, 94 are PWF, and 95 are OSF – note that some machine defects are a combination of multiple failures. The imbalance in the data can be seen as an indication for up-sampling approaches such as SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Still, our experiments showed no significant improvement, and our reported models are trained on the provided data only. We performed experiments using five different model architectures implemented by scikit-learn[1]: k-nearest neighbors , Decision Tree, Random Forest, Gradient Boosting, and Neural Network, and report the results in Table 1, as well as the breakdown of the best-performing model in Table 2, where we see that the detection of the HDF and PWF are more trivial than ODF, which is consistent among the models. Therefore, we can assume that the misclassification of machine failure is potentially caused by detecting the OSF.

4.2. Parameter-Setting Recommendations

MOXAI uses DiCE as an explainer backend, which was initialized using the trained model and the training data. We use the genetic algorithm provided by DiCE, as it supports parameters that prioritize counterfactuals similar to training data and thus avoid regions in the parameter space that are not well defined due to missing training data. We allow variation in all features, but real-life use cases will likely require limiting the parameters that can be modified at the machine.

[1] https://scikit-learn.org

Table 3. Accuracy of suggested parameters.

	KNN	DT	RF	GBM	MLP
HDF	0.89	1.0	1.0	0.94	0.78
PWF	0.91	1.0	1.0	1.0	0.71
OSF	0.58	0.95	0.95	0.84	0.47
overall_accuracy	0.81	0.98	0.98	0.93	0.68

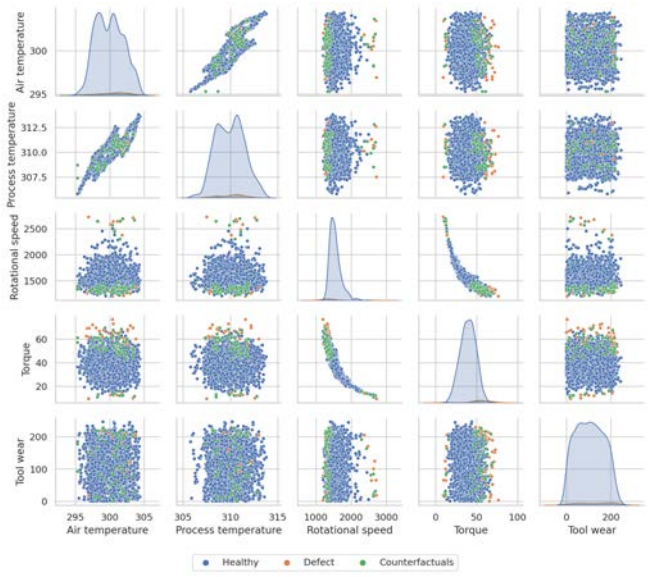


Figure 3. Pairplot of the test set and created counterfactual samples, for the five different numerical features characterizing the milling process.

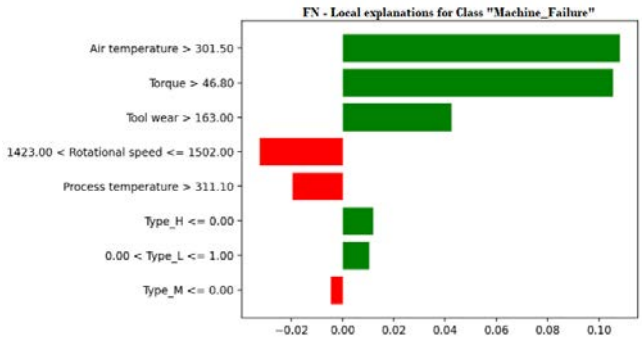


Figure 4. LIME's local explanations for a misclassified sample as not a machine failure (FN).

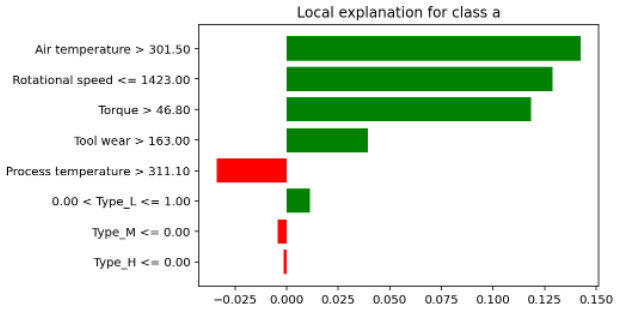


Figure 5. LIME's local explanations for a sample correctly classified as a machine failure (TP).

To evaluate MOXAI, we use the analytic definition of errors provided by the AI4I dataset creators. For each defective sample of the test set, we use MOXAI to calculate a suggested set of machine parameters. Figures 2 and 3 depict the healthy and defective samples in the test set and the generated counterfactuals. We take the error definition to determine if the solution proposed by MOXAI actually solved the problem and corresponded to a healthy product. We report the percentage of successful corrections as accuracy in Table 3.

4.3. Correction of Model Failures

We generate LIME explanations for each failed sample using the models discussed in the preceding section. We encounter 21 failed samples, comprising 15 false negatives (FNs) and six false positives (FPs). Through the analysis of modeling separated failure modes, we noticed that these misclassifications predominantly stem from the model's failed attempt to detect PFW and OSF accurately. We extract the explanations using the algorithm detailed in Section 3.2. To correct false positives (FPs), we randomly generate float values within the approximate feature range identified by LIME to produce counterfactual instances. We leverage our understanding of value ranges given by dataset providers, contributing to failures in each specific mode, to rectify errors in parameter settings. Similarly, we apply this method to false negatives (FNs), mostly from inaccuracies in process temperature values. By generating counterfactual instances, we illustrate the adjustments required in parameter values to identify defective samples accurately. In Figure 4 and Figure 5, we demonstrate a comparison of a true positive (correctly detected machine failure) with a scenario where the model predicted a failure as "not a failure" with low confidence (the prediction probability for the class Failure is 0.47) explained by LIME. The plot shows that, even though the features *Air temperature, Torque* and *Tool wear* are positively contributing to this prediction being a failed sample, the values of *Rotational speed* and *Process temperature* are shifting the model's decision towards the class "not a Failure". MOXAI suggests a minor change of *Process Temperature* to a value slightly smaller than 311.10, creates a counterfactual, and corrects this prediction. In practice, the domain expert should verify whether this change is valid and does not contradict the definition of this failure mode.

5. Discussion

The evaluation of model performance in a classification task unveils nuanced observations. While all models exhibit satisfactory accuracy in data classification, their effectiveness

in generating reliable counterfactual samples varies. Notably, tree-based models emerge as the most robust, surpassing alternative methodologies, such as Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN). Furthermore, an analysis of error types reveals differences among model performances. Specifically, most models demonstrate proficiency in addressing tool wear (PWF) and heat dissipation (HDF) errors but struggle when confronted with errors arising from multiple product types (OSF). These findings underscore the importance of assessing classification accuracy and considering models' ability to provide dependable counterfactual samples and their efficacy in handling diverse error types. Moreover, the current state of MOXAI is limited to the parameters within the proximity of its training set. Therefore, it cannot suggest optimizations for unseen production scenarios. One approach to address this limitation could be the usage of digital twin solutions that are more flexible when it comes to approximating new parameters and production settings.

We underscore the significance of comprehending the model's decision-making process in failure detection and why these particular counterfactuals were suggested. A guided approach aids in understanding why a model failed and whether the model's identified correlations are logical. While MOXAI offers an interpretable and human-in-the-loop system for comprehending model failures and suggesting meaningful samples tailored to this specific use case, the semi-automatic counterfactuals produced by our human-guided approach could benefit from considering feature co-linearities and interactions, and a domain expert should verify them to exclude nonsensical examples. This process is crucial for gauging the model's reliability and assessing the suitability of a fully automated counterfactual generation module. Therefore, the operator can plainly trust the model's recommendations to choose the best settings based on the explanations provided by LIME's output.

6. Conclusion

The approach of XAI to enhance data-driven process tuning for optimizing production or predictive maintenance is promising. MOXAI proposes a data-driven, XAI-powered approach to optimizing manufacturing machine parameters, relying on pre-trained ML models of any nature. We have trained different ML models for failure prediction in a popular synthetic dataset representing a realistic industrial scenario, applying MOXAI's information flow to identify potential corrections to improve failure samples and improve understanding of the operation of these ML models.

DiCE is a key element in automatically generating counterfactual explanations, which can assist operators in adjusting process settings so that machine failures or degraded production quality can be reduced. Applying LIME explanations to address false predictions within our model proved insightful. We successfully rectified both false positives and false negatives by analyzing failure modes and generating counterfactual instances based on LIME insights. Additionally, our demonstration of LIME's output underscores its potential to enhance model decisions' interpretability.

Enhancing the understanding of counterfactual methods is important for future advancements. This ensures that such methods foster a causal understanding for human operators while avoiding any risks of biased, sub-optimal, or erroneous explanations.

Acknowledgments

This work has been funded by the Austrian Research Promotion Agency (FFG) under grant No. 883864 *ZDM – Zero Defect Manufacturing* and grant No. FO999913202 *UNDERPIN* and by the European Commission under contract No. 101123179 *UNDERPIN*.

References

Ameli, M., Becker, P. A., Lankers, K., van Ackeren, M., Bähring, H., & Maaß, W. (2022). Explainable unsupervised multi-sensor industrial anomaly detection and categorization. In *2022 21st ieee international conference on machine learning and applications (icmla)* (pp. 1468–1475).

Ates, E., Aksar, B., Leung, V. J., & Coskun, A. K. (2021). Counterfactual explanations for multivariate time series. In *2021 international conference on applied artificial intelligence (icapai)* (pp. 1–8).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Jakubowski, J., Stanisz, P., Bobek, S., & Nalepa, G. J. (2021). Explainable anomaly detection for hot-rolling industrial process. In *2021 ieee 8th international conference on data science and advanced analytics (dsaa)* (pp. 1–10).

Jakubowski, J., Stanisz, P., Bobek, S., & Nalepa, G. J. (2022). Roll wear prediction in strip cold rolling with physics-informed autoencoder and counterfactual explanations. In *2022 ieee 9th international conference on data science and advanced analytics (dsaa)* (p. 1-10). doi: 10.1109/DSAA54385.2022.10032357

Jalali, A., Haslhofer, B., Kriglstein, S., & Rauber, A. (2023). Predictability and comprehensibility in post-hoc xai methods: A user-centered analysis. In *Science and in-*

formation conference (pp. 712–733).

Kulesza, A., Taskar, B., et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, *5*(2–3), 123–286.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319-330. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446 doi: https://doi.org/10.1002/asmb.446

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Matzka, S. (2020). Explainable artificial intelligence for predictive maintenance applications. In *2020 third international conference on artificial intelligence for industries (ai4i)* (pp. 69–74).

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (p. 607–617). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3351095.3372850 doi: 10.1145/3351095.3372850

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 1135–1144). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2939672.2939778 doi: 10.1145/2939672.2939778

Schockaert, C., Macher, V., & Schmitz, A. (2020). Vae-lime: deep generative model based approach for local data-driven model interpretability applied to the ironmaking industry. *arXiv preprint arXiv:2007.10256*.

Seiffer, C., Ziekow, H., Schreier, U., & Gerling, A. (2021). Detection of concept drift in manufacturing data with shap values to improve error prediction. In *Data analytics 2021: The tenth international conference on data analytics* (pp. 51–60).

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*, *68*(8), 5704–5723.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th international conference on machine learning - volume 70* (p. 3145–3153). JMLR.org.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2921–2929).

NLP-Based Fault Detection Method for Multifunction Logging-While-Drilling Services

Corina Maria Panait[1], Qian Su[2], Nahieli Vasquez[3], Ahmed Mosallam[3], Fares Ben Youssef[2], Olexiy Kyrgyzov[3], Hassan Mansoor[4], Anup Arun Yadav[5]

[1] *SLB, Str. Sergeant Constantin Ghercu, Bucharest, Romania*
CPanait5@slb.com
[2] *SLB, 135 Rousseau Road, 70592 Youngsville, Louisiana, USA*
QSu3@slb.com
FYoussef@slb.com
[3] *SLB, 1 rue Henri Becquerel, 92140 Clamart, France*
NVasquez23@slb.com
AMosallam@slb.com
OKyrgyzov@slb.com
[4] *SLB, Nowogrodzka Street 68, 02-014 Warsaw, Poland*
HMansoor2@slb.com
[5] *SLB, 8 Office 301 Commerzone IT park, Pune, India*
AYadav28@slb.com

ABSTRACT

This paper presents a Natural Language Processing (NLP) method aimed at detecting faults within field failure reports of drilling tools. It builds on the definition of entities specifically matched to our unique requirements. These entities have been annotated within the dataset under the guidance of a Subject Matter Expert (SME), laying a foundation for our NLP method. By utilizing a model based on bidirectional encoder representations from transformers, the method achieves an F1-score of 88% in identifying entities and consequently detecting faults within field failure reports. This work is part of a long-term project aiming to construct a failure analysis and resolution system for drilling tools.

1. INTRODUCTION

The oil and gas industry relies heavily on logging tools that operate in extreme environmental conditions, including elevated temperatures, vibrations, and pressures. Such conditions can accelerate the degradation of tools, leading to potential failures. These failures not only compromise operations by providing inaccurate information but also result in delayed deliverables, tool repair, or even cancellation of the entire operation. Such setbacks translate into nonproductive time and substantial financial losses. Efficiency and speed in the maintenance process are critical when a logging tool fails. The maintenance team tackles the task of navigating extensive unstructured data to identify patterns of failures. Streamlining this maintenance workflow is essential to expedite the turnaround time of the tool, achieving its swift return to operational readiness and minimizing nonproductive intervals and associated financial losses. Manual analysis of these data proves extremely time-consuming in the context of an industry where avoiding downtime is a priority.

While Prognostics and Health Management (PHM) methods have been successfully utilized to enable predictive maintenance, this approach has its limitations (Mosallam, Laval, Youssef, Fulton, & Viassolo, 2018; Mosallam, Kang, Youssef, Laval, & Fulton, 2023; Mosallam, Youssef, et al., 2023; Kang et al., 2022). It primarily relies on equipment sensor data and does not take into account the rich source of information in maintenance logs, such as failure reports, asset performance, maintenance policies, and failure patterns collected during the life cycle of asset management. Through computerized analysis, i.e., NLP, it is possible to make the process considerably more efficient (Stenström, Al-Jumaili, & Parida, 2015). For instance, (Juan Pablo Usuga Cadavid & Fortin, 2020) utilized an NLP model named CamemBERT (Bidirectional Encoder Representations from Transformers) to predict the criticality and duration of maintenance issues based on free-form text comments from operators (Martin et al., 2019). Despite the unstructured and imbalanced nature of maintenance logs, the authors suggest that their approach can pro-

vide significant advantages in coupling production scheduling with maintenance logs, enabling the adaptation of planning to the shop floor. Despite the recent and rapid development of the NLP field, extracting meaningful insights from maintenance logs remains a challenging task. (Brundage, Sexton, Hodkiewicz, Dima, & Lukens, 2021) argue that current NLP tools are not suitable for engineering data and propose a domain-driven approach called Technical Language Processing (TLP). They suggest that key NLP tools need to be adapted to the maintenance domain based on available maintenance text-based data. For example, (Naqvi, Ghufran, et al., 2022) introduced a TLP approach utilizing a Case-Based Reasoning (CBR) framework paired with a domain-adapted BERT model to address maintenance issues through textual data from mining operations. This approach, particularly the use of a Transformer-based Sequential Denoising Autoencoder (TSDAE) for unsupervised fine-tuning and cosine similarity for case assessment, underscored the importance of domain-specific model training (Wang, Reimers, & Gurevych, 2021). (Lee & Marlot, 2023) proposed Oil & Gas domain-relevant entity and relationship extraction from drilling reports. The approach involves training a Named Entity Recognition (NER) model to identify key information or failure symptoms in the reports, such as equipment, operations, or events, which can be considered a fault detection task from maintenance logs. This is followed by a Relation Extraction model that identifies the relationships between the entities and recommends early mitigation using historical data. The authors also applied data augmentation techniques to increase the data samples and improve the model's performance in detecting rare entities. Finally, (Naqvi, Varnier, Nicod, Zerhouni, & Ghufran, 2022) propose an NLP method for fault diagnostics from maintenance logs. The study finds that fine-tuning CamemBERT outperforms classical NLP approaches and that data augmentation using deep contextualized embedding further improves performance.

We present an NLP-based fault detection approach leveraging a NER model based on the BERT framework for Logging-While-Drilling Service (LWD) (Hansen & White, 1991) (Figure 1). This approach addresses equipment failures within unstructured data, offering a robust solution for the oil and gas industry. The BERT model, chosen for its advanced contextual understanding of words in text, excels at identifying and classifying key entities related to equipment failure and operational procedures within textual data. Custom and actionable entities were defined through extensive data labeling to enhance the model's proficiency in recognizing relevant information.

The rest of this paper is structured into four sections. Section 2 presents a description of the failure investigation process. The method and the results are presented in section 3 and 4, respectively. Finally, section 5 concludes the paper.

Figure 1. Multifunction LWD Service.

2. Failure Investigation Process

Field incidents and maintenance data are utilized to construct the dataset required for this work. These data are stored across various business systems in diverse formats. Field incidents are reported in the Failure Investigation Business System known (FIBS). Field Crew generates a field failure report which includes the following details:

1. Basic event information, including event date, location and the suspected failed technology.
2. The primary content is the "Field Failure Description," a free-text input where the Field Crew documents all their observations concerning the event sequence and the failure symptoms of suspected technology. This part is the cornerstone of our analysis, as identifying historical events with similar failure symptoms is crucial.
3. The "Remedial Actions Attempted," also a free-text input, where the Field Crew records all the actions that they attempted to restore the tool to normal operation.

Maintenance Data is primarily stored in a maintenance business system. A field failure Work Order (WO) is generated for each technology implicated in field incidents and flagged by the field crew when creating the FIBS report. The main section of each Work Order for technicians involved in failure investigations is the failure description section, where they provide free-text input on analysis, testing, and findings. This segment serves as a valuable source of insights and knowledge, helping others who may encounter similar failures with the same technology. After completing the investigation and if any failed components are identified, this failed part is then recorded in a dedicated business system. The failed components are confirmed immediate causes in all historic incidents, which can be utilized to establish a failure Pareto chart, illustrating the probability of specific causes for distinct failure symptoms. The failure investigation process is summarized in Figure 2.

Figure 2. Failure Investigation Process.

A critical part of this process is analyzing failure reports of LWD service to determine whether the downhole conditions

have had any detrimental effects on the operability. Due to the complexity of this service, analysis of this vast amount of data is very time-consuming and prone to error if performed manually. Therefore, fault detection from maintenance logs is of utmost importance for operation. An automated tool, which can determine different failure symptoms from maintenance logs with minimal user input, removes variability, eliminates human error and provides an efficient decision on the required maintenance in a fraction of the time. The reliability benefits are clear and provide significant cost savings both for the client in terms of reduced Non-Productive Time at the rigsite and for the Original Equipment Manufacturer in terms of reduced Materials and Supplies (M&S) during maintenance and troubleshooting.

3. Proposed Method

Currently, this work focuses solely on one technology, a multifunction Logging-While-Drilling tool designed for oil and gas well drilling applications. This technology integrates a comprehensive suite of formation-evaluation measurements (resistivity, porosity, density, natural gamma-ray, etc.) and drilling parameters (temperature, pressure, shock, vibration, etc.) into a single housing. Given the domain-specific nature of the data, it was necessary to create bespoke entities tailored to our unique requirements. This crucial task was carried out by an SME. The SME annotated key entities within a vetted dataset, laying the groundwork for a robust NLP model. This model is adept at identifying and categorizing phrases that fall into predefined entity groups, each critical for deciphering the complex narratives within the data:

1. Failure Symptom: This category captures the explicit details of failures, such as 'png stopped firing' or 'no porosity data.' Identifying these allows for a precise understanding of the failure characteristics.

2. Data Channels: These entities encompass technical log parameter values like 'state changing to 2304.' Their recognition is vital for correlating technical readings with failure events.

3. Operational Actions: This group includes actions taken in response to issues, such as 'downlink to shutdown png.' Understanding these actions aids in assessing the effectiveness of operational responses.

4. Drilling Conditions: These entities report on the physical conditions during drilling, like 'top cement was tagged.' Recognizing these conditions is essential for contextualizing failures within their operational environment.

The proposed method consists of three steps: data collection, data preparation, and modelling, as illustrated in Figure 3.

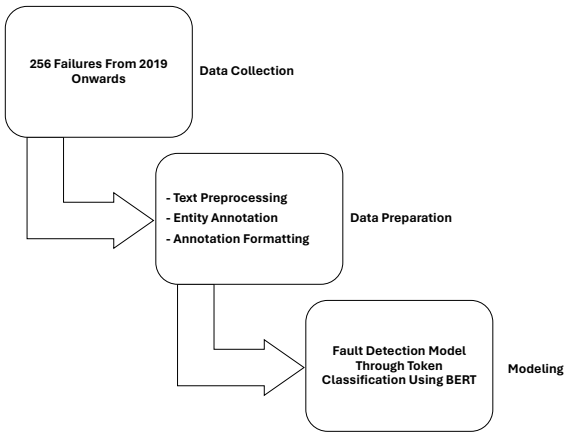


Figure 3. Proposed Method.

3.1. Data Collection

The dataset comprises 256 failure descriptions, extracted from internal database tables, focusing on data from the year 2019 onwards. This time restriction was applied because the majority of relevant and validated data, as identified by the SME, begins from this period.

3.2. Data Preparation

The data processing involved three main steps: text preprocessing, entity annotation, and adaptation to the dataset format of the (*Hugging Face*, Accessed: 2024-05-27) library, a widely-used platform that provides pre-trained models for natural language processing tasks:

1. Text Preprocessing: In this step we concentrated on removing non-alphanumeric characters and converting all text to lowercase to ensure uniformity and reduce complexity in the dataset. Additionally, we removed sections of text that originated from application-dependent formatting, such as incident dates and job numbers, as these did not contribute valuable information for fault detection.
2. Entity Annotation: We utilized syntactic strategies to ensure qualitative consistency throughout the entire dataset. A significant decision in this process was to include verbs and adverbs in all entity categories. This approach was adopted to maintain grammatical consistency across entities, which is crucial for reducing the risk of misclassification. The annotation was executed by the SME using the Doccano application, resulting in a JSON (JavaScript Object Notation) file, a format used for storing structured data. This file contains failure descriptions and lists of three chained elements, detailing the starting character, ending character, and associated category for each entity. While only 256 reports were labeled, the annotated data itself is extensive, with each failure description often containing multiple annotated phrases and entities.

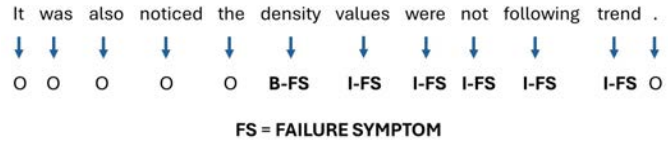


Figure 4. Hugging Face IOB Formatting.

3. Annotation Formatting: We converted generated annotations into a format compatible with the Hugging Face token classification paradigm. After validating the model, a corresponding tokenizer was utilized to divide the text into subwords. Subsequently, the character spans from Doccano were converted into Inside-Outside-Beginning tagging (IOB) format, a requirement for Hugging Face. An example of such formatting can be observed in Figure 4, where some tokens in a sentence are assigned the corresponding IOB formatted *FAILURE SYMTPOM* labels. The total amount of BI tags is $11,333$ as illustrated in Figure 5. The tokens ids and attention masks were also retained during the final conversion to the Dataset object. This object adheres to best practices for data splitting in cross-validation, featuring a distribution of 204 instances for training, 26 for validation, and 26 for testing.

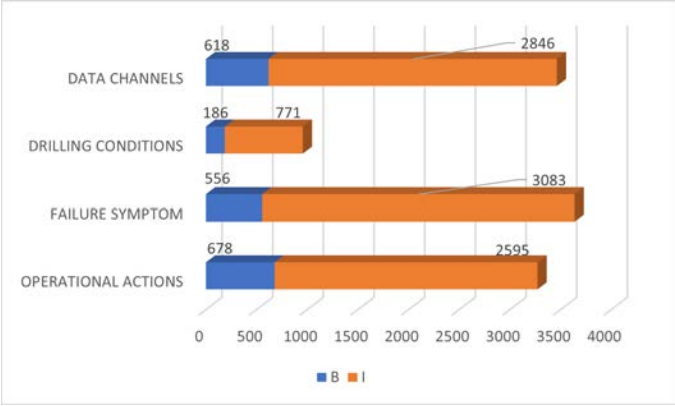


Figure 5. Bar Distribution Graph of BI-tags.

3.3. Modeling

Formally, we can define every failure description report, X, as a sequence of tokens:

$$X = \{x_0, x_1, ..., x_n\} \tag{1}$$

where x_i represents the i_{th} token in the report.

Our objective is to obtain a sequence of predicted labels, Y,

$$Y = \{y_0, y_1, ..., y_n\} \tag{2}$$

where y_i is the label of the x_i token.

For this purpose, we began by examining various unsupervised learning models to extract insights from text. However, we found that at the time of solution development, none were trained on an industry-specific corpus. Most available NER models were designed to identify general entities like date, location, person, and company, which did not align with our business-specific needs.

After completing the annotation of entities, we undertook a comparative analysis of several NLP models. This included, but was not limited to, Spacy NER, "distilbert-base-uncased", and "bert-base-cased". Our objective was to identify the most suitable model for our fault detection use case. Among these, "bert-base-cased" demonstrated initially promising results. Consequently, we focused on fine-tuning its hyperparameters to optimize performance.

The BERT model is pretrained on two tasks: masked language modeling and next sentence prediction. Each token in BERT is represented by a combination of its token embedding, segment embedding, and position embedding. During our fine-tuning process, we utilized BERT's default activation function, GELU, along with a final classification layer that employs a softmax function to determine class probabilities. For token classification, we extract the hidden layer representation of each token and apply the softmax function to compute the probabilities for each class.

The formula for the softmax function is detailed below, where K is the number of classes and z_i is the output of the classification layer:

$$Z = [z_0, z_1, ..., z_K] \tag{3}$$

$$s(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4}$$

4. Results

In line with the time constraints of our business objectives, we explored the hyperparameter space for the "bert-base-cased" model exhaustively. To evaluate our model's performance, we primarily focused on the F1-score, a metric derived from precision and recall. The formulas for precision, recall, and F1-score are provided below:

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - score = 2(\frac{precision * recall}{precision + recall}) \tag{7}$$

Entity	Precision	Recall	F1-score	No. of Instances
B-DATA CHANNELS	0.92	0.94	0.93	3816
I-DATA CHANNELS	0.81	0.74	0.78	2378
B-DRILLING CONDITIONS	0.50	0.18	0.27	227
I-DRILLING CONDITIONS	0.59	0.27	0.37	1073
B-FAILURE SYMPTOM	0.63	0.80	0.71	450
I-FAILURE SYMPTOM	0.72	0.79	0.75	2711
B-OPERATIONAL ACTIONS	0.50	0.78	0.61	460
I-OPERATIONAL ACTIONS	0.59	0.49	0.54	1708
macro avg	0.69	0.64	0.68	4762
weighted avg	0.88	0.88	0.88	4762

Table 1. Classification Report of BERT Checkpoint.

This effort led to a satisfactory model checkpoint, achieving an F1-score of 88%. This score reflects the model's precision and reliability in recognizing entities (See Table 1).

The most effective entities identified by the model were Data Channels and Failure Symptom, with weighted averages of 81% and 74% respectively. These averages were calculated by multiplying the F1-score by the corresponding support value and dividing by the sum of support for each BI-tags pair. Operational Actions showed average performance with a 55% score, while Drilling Conditions lagged at 35%. The lower performance in these categories correlates with a reduced token support count, indicating the model's sensitivity to class imbalances.

Our analysis of the text data revealed that the weakly predicted categories had less variance in our validated data sample, underscoring the need for more representative data for these categories. Other insights pertain to the nature of technology failures. We observed that there are a limited number of ways in which failures manifest, and this limitation aided the model's performance in identifying well-represented entities. Specifically, certain phrases indicating a specific techlog parameter value or failure symptom recur more frequently in the failure descriptions.

5. Conclusions

In this paper, we presented a fault detection NLP-based method from maintenance logs. The methods builds on identifying four technical-defined entities, essential to the failure investigation process. This approach entailed fine-tuning "bert-base-cased" model which achieved an F-1 score of 88%, underscoring the model's precision and reliability in recognizing critical entities.

The practical implications of our work are significant, with the potential to improve operational decision-making through enhanced pattern recognition in historical failure data. The impact of our findings is geared towards improving operational efficiency, reducing downtime, and cutting costs.

In future work, we aim to expand our research from identifying failures to comprehensively diagnosing them. This will involve a more detailed examination of failure events to extract insights into their causes and impacts. By advancing from simple detection to in-depth diagnostics, we will offer not just identification but also solutions.

Additionally, we intend to develop a Case-Based Reasoning system that will complement our NLP framework. This system will feature a similarity model to gather and compare similar cases, bringing forward solutions that have been effective in the past. This enhancement is expected to not only pinpoint failures but also recommend validated resolutions, thereby streamlining the path from problem recognition to problem-solving.

The integration of the CBR system is expected to leverage historical insights and expert knowledge, evolving into a dynamic model that improves with each new dataset. This step will mark a significant advance in intelligent fault detection systems, pushing the boundaries of what is currently possible in operational efficiency and safety.

References

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, *27*, 42-46. doi: https://doi.org/10.1016/j.mfglet.2020.11.001

Hansen, R., & White, J. (1991). Features of logging-while-drilling (lwd) in horizontal wells. In *Spe/iadc drilling conferencespe/iadc drilling conference.* doi: 10.2118/21989-MS

Hugging Face. (Accessed: 2024-05-27). `https://huggingface.co/`.

Juan Pablo Usuga Cadavid, S. L. R. P., Bernard Grabot, & Fortin, A. (2020). Valuing free-form text data from maintenance logs through transfer learning with camembert. *Enterprise Information Systems*, *16*(6), 1790043. doi: 10.1080/17517575.2020.1790043

Kang, J., Varnier, C., Mosallam, A., Zerhouni, N., Youssef, F. B., & Shen, N. (2022). Risk level estimation for electronics boards in drilling and measurement

tools based on the hidden markov model. In *2022 prognostics and health management conference (phm-2022 london)* (p. 495-500). doi: 10.1109/PHM2022-London52454.2022.00093

Lee, M., & Marlot, M. (2023). Information Retrieval from Oil and Gas Unstructured Data with Contextualized Framework. , *2023*(1), 1-5. doi: https://doi.org/10.3997/2214-4609.202332039

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., ... Sagot, B. (2019). Camembert: a tasty french language model. *CoRR*, *abs/1911.03894*. Retrieved from http://arxiv.org/abs/1911.03894

Mosallam, A., Kang, J., Youssef, F. B., Laval, L., & Fulton, J. (2023). Data-driven fault diagnostics for neutron generator systems in multifunction logging-while-drilling service. In *2023 prognostics and health management conference.*

Mosallam, A., Laval, L., Youssef, F. B., Fulton, J., & Viassolo, D. (2018). Data-driven fault detection for neutron generator subsystem in multifunction logging-while-drilling service. In *PHM society european conference.*

Mosallam, A., Youssef, F. B., Sobczak-Oramus, K., Kang, J., Gupta, V., Shen, N., & Laval, L. (2023). Data-driven degradation modeling approach for neutron generators in multifunction logging-while-drilling service. In *2023 prognostics and health management conference.*

Naqvi, S. M. R., Ghufran, M., Meraghni, S., Varnier, C., Nicod, J.-M., & Zerhouni, N. (2022). Human knowledge centered maintenance decision support in digital twin environment. *Journal of Manufacturing Systems*, *65*, 528-537. doi: https://doi.org/10.1016/j.jmsy.2022.10.003

Naqvi, S. M. R., Varnier, C., Nicod, J.-M., Zerhouni, N., & Ghufran, M. (2022). "Leveraging Free-Form Text in Maintenance Logs Through BERT Transfer Learning". In L. Troiano, A. Vaccaro, N. Kesswani, I. Díaz Rodriguez, & I. Brigui (Eds.), *Progresses in artificial intelligence & robotics: Algorithms & applications* (pp. 63–75). Cham: Springer International Publishing.

Stenström, C., Al-Jumaili, M., & Parida, A. (2015). Natural language processing of maintenance records data. In *International journal of comadem* (Vol. 18).

Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *CoRR*, *abs/2104.06979*. Retrieved from https://arxiv.org/abs/2104.06979

BIOGRAPHIES

Corina Maria Panait is a Data Scientist at SLB Romania, within the Data Science & AI Hubs. Her 1.5 years of experience in the company concentrate on NLP solutions, designed to obtain quick insights from unstructured text data in the following business areas: PHM, Health & Safety and Generative AI. Corina holds a Master's in Data Science from the University of Bucharest and a Bachelor's in Economics Cybernetics from the Bucharest Academy of Economic Studies.

Qian Su is currently a Technical Engineer for the Offshore Atlantic Basin Well Construction. In her role, she provides technical support with a focus on electronics for failure investigations, conducts reliability improvement initiatives, and digital applications. Qian holds a Master's degree in Electrical Engineering from the Institut National des Sciences Appliquées de Lyon.

Nahieli Vasquez is a Data Scientist with SLB France for around a year and a half. She has several years of experience working in NLP applications. Prior to this position, she has worked in applying NLP methods to curriculum vitae scoring and social network moderation. Her areas of interest include NLP, Gen AI, Cloud services and End-to-end product development. She holds a Master Degree in Quantitative Economics and Econometrics from El Colegio de Mexico.

Ahmed Mosallam is the Data Science & AI European Hub Manager at SLB technology center in Clamart, France. He has his Ph.D. degree in automatic control in the field of PHM from University of Franche-Comte in Besanc¸on, France. His main research interests are signal processing, data mining, machine learning and PHM.

Fares Ben Youssef Fares Ben Youssef is the Reliability and COSD Manager, where he leads a team of Engineers and Technicians focused on several aspects of Well Construction for Offshore Atlantic basin. His responsibilities include delivering Fit for Basin projects, which involve digital, material, mechan-

ical, and electrical design changes tailored to the Offshore Atlantic basin requirements. In 2011, Fares earned a Master's degree in Electronics and Telecommunication Engineering from the University of Paris-Saclay, France.

Olexiy Kyrgyzov graduated with B.S. and M.S. in CS in 2001 and 2002, respectively, from Dnipro National University (Dnipro, Ukraine). He received his Ph.D. in EE from the Northeastern University (Boston, USA) in 2010, where he worked on research projects as an assistant of Prof. Deniz Erdogmus. The topic of his dissertation was 'Non-redundant tensor decomposition'. After graduation he moved to France and worked on a wide range of projects in state research centers. He is currently a senior data scientist in DS&AI Hub Eur at SLB (Clamart, France). His research focuses on machine learning and artificial intelligence with applications to signal processing, natural language processing, and image analysis. He works to solve real-world Oil & Gas data analysis problems with delivering final compact and efficient products.

Hassan Mansoor is Full Stack Developer at SLB Poland. He holds a Master's Degree in Software Engineering from Poznan University of Technology, Poland. He has over 7 years of experience as a software developer, where he has worked with several cloud and web technologies.

Anup Arun Yadav is an AI Application Manager at SLB, PITC. He holds a Master's Degree in Computer Science from Pune University, India. He has over 13 years of experience as an application developer, where he has worked with several cloud and web technologies.

Noise-aware AI methods for robust acoustic monitoring of bearings in industrial machines

Kerem Eryılmaz[1], Fernando de la Hucha Arce[2], Jeroen Zegers[3], and Ted Ooijevaar[4]

[1, 3, 4] *Flanders Make, Gaston Geenslaan 8, 3001 Leuven, Belgium*
kerem.eryilmaz@flandersmake.be
ted.ooijevaar@flandersmake.be

[2] *Flanders Make, Graaf Karel De Goedelaan 16-18, 8500 Kortrijk, Belgium*
fernando.delahuchaarce@flandersmake.be

ABSTRACT

Traditionally, companies have relied on vibration based condition monitoring technologies to implement condition based maintenance strategies. However, these technologies have drawbacks, such as the requirement of contact accelerometers. As an alternative, acoustic condition monitoring is non-invasive and allows for easy deployment. Furthermore, the use of microphones potentially enables the monitoring of multiple components using a single sensor, making the monitoring system scale better with machine or production complexity. However, microphone signals typically show a low signal-to-noise ratio (SNR), impacted by the high level of background noise which is often present in industrial environments. Particularly, the traditional method for monitoring the health condition of rolling element bearings, based on assessing whether the squared envelope spectrum of the bearing signal exceeds a given threshold at the fault frequencies, cause too many false positives when applied directly to microphone signals. It is therefore crucial to develop strategies to increase the robustness of acoustic monitoring methods. In this paper, we present and evaluate two data-driven strategies to robustly diagnose bearing faults from a microphone signal. Our proposed strategies are noise weighting based on the detection of background noise, and an artificial intelligence (AI) model that uses as input a combination of the traditional bearing fault frequencies and the mel spectrum of the microphone signal. These methods leverage both domain knowledge and data-driven techniques to increase the detection robustness. Our approach is implemented as a model trained and tested on bearing accelerated lifetime tests performed in the Smart Maintenance Lab setup at Flanders Make. Our results show that the use of our proposed strategies leads to significant improvements in diagnostic performance and time to first detection over noise-unaware acoustic monitoring methods.

1. INTRODUCTION

Condition monitoring involves the continuous monitoring of machine parameters to detect changes that are indicative of a developing fault. This a key component of condition based maintenance, the strategy that schedules maintenance actions based on the current health diagnosis of machine components, with the goal of reducing equipment downtime and total maintenance cost. The detection of faults in rolling element bearings is of special interest, since they are critical components of rotating machinery, and their faulty signals are often masked under other dominant sources (Randall, 2011). The use of accelerometers is the most common approach for monitoring bearing and gear faults, as vibrations often carry early information of their incipient damages (Lee et al., 2014).

There exist a wide range of well-established signal processing methods that are applied to vibration signals in order to estimate the health condition of a bearing. One of the most successful methods is envelope analysis, whose comprehensive description is given in (Randall & Antoni, 2011). It relies on the extraction and tracking of the fault characteristic frequencies in the squared envelope spectrum (SES) of the vibration signal generated by the bearing. As their name suggests, these frequencies are related to the bearing faults, and contain an increasing amount of energy as faults become more serious. For bearings operating under conditions of low load and low rotational speed, a different method based on stochastic resonance is proposed and shown to outperform envelope analysis in (Ompusunggu, Devos, & Petre, 2013).

However, a disadvantage of diagnosis techniques based on vibration analysis is that the accelerometers should be mounted close to the rotating component of interest. Consequently, several accelerometers are needed to monitor multiple bear-

ings, and accessibility constraints may render this impossible.

As an alternative to overcome these limitations, acoustic non-contact sensors such as microphones have recently drawn attention mainly for two reasons. First, they allow for easier deployment, as they do not need to be physically mounted by bolts, glue or magnets. Second, microphone signals may acquire information from several bearing signatures, potentially enabling the monitoring of multiple components with fewer sensors than in schemes based on accelerometers. Nevertheless, microphones will unavoidably collect signals from undesirable noise sources mixed with the signals emitted by the bearings. In industrial environments, these noise sources are generally quite strong and varied in nature, leading to microphone signals with low signal-to-noise ratio (SNR) that result in poor diagnosis performance. For this reason, dedicated methods to increase the robustness to background noise are crucial in acoustic monitoring.

Due to the wide array of different noise sources present in industrial environments, data-driven strategies are a powerful tool to increase the robustness of acoustic monitoring. In a data-driven strategy, healthy and damaged bearings are classified by a data-driven model trained using a set of relevant acoustics features. In (Mian, Choudhary, & Fatima, 2022), six sound quality features from microphone signals were used to train a support vector machine to diagnose bearing damages. For the diagnosis of bearing, rotor and stator faults in induction motors, a frequency domain feature extractor method combined with a nearest neighbour classifier is proposed and shown to perform well in (Glowacz, 2019).

Another commonly used strategy to achieve robust acoustic monitoring relies on microphone arrays and beamforming. The works presented in (Cardenas Cabada, Leclere, Antoni, & Hamzaoui, 2017; Ricardo Mauricio, Denayer, & Gryllias, 2022, 2023), and references therein, show that this strategy can produce good diagnosis results for bearing monitoring using beamforming. However, the requirement of multiple microphones and precise positioning increases the practical complexity of implementing this solution. For this reason, we consider beamforming strategies outside of the scope of this work, and focus on data-driven strategies using a single microphone.

In this paper we propose two data-driven methods to increase the robustness of the diagnosis of bearing faults using acoustic sensing. Our first approach is noise weighting based on the detection of background noise, and the second one is an artificial intelligence (AI) model whose input is a combination of the bearing fault frequencies and the mel-spectrum of the microphone signal. These methods integrate both domain knowledge and a data-driven technique, and they are trained and tested on bearing accelerated lifetime experiments performed in the Smart Maintenance Lab setup at Flanders Make. The goal is to evaluate the performance of our proposed methods, and show that acoustic monitoring is a cost effective and practical alternative to vibration monitoring.

The rest of this paper is structured as follows. In Section 2, the well-established envelope analysis method for bearing fault diagnosis is reviewed, and an explanation of its poor performance when applied to acoustic sensing is provided. Our two proposed data-driven methods for robust acoustic bearing fault diagnosis are detailed in Section 3. A description of the experimental setup is given in Section 4, which includes the performed bearing accelerated lifetime experiments, the acoustic scene, and the parameters of the signal processing and AI models. The performance of our proposed methods is evaluated and discussed in Section 5. Finally, the main conclusions are summarized in Section 6.

2. PROBLEM STATEMENT

Rolling element bearings are a crucial component in a wide variety of rotating machinery. However, over time they can develop faults such as surface fatigue defects or wear. For localized faults, as the rolling elements strike a fault in the inner or outer race, an impulse is generated that excites high frequency resonances on the structure between the bearing and the sensor location.

2.1. Vibration-based bearing fault diagnosis

The vibration signals from a faulty bearing can be modelled as a modulated blend of several signal components: an impulsive signal associated with the fault, the high frequency signals related to the the dynamics of other machine components such as the shaft and gears, the modulation between these signals and additional noise. The well-established method for bearing diagnostics is the so-called envelope method, which first enhances the impulsive signal generated by the fault, and then estimates the energy at the fault characteristic frequency and its harmonics from its squared envelope spectrum (SES). A complete explanation of the method is provided in (Randall & Antoni, 2011).

In this paper we focus on inner race faults, for which the fault characteristic frequency is the ball pass frequency, inner race (BPFI), given by

$$f_{\text{BPFI}} = \frac{n f_r}{2} \left\{ 1 + \frac{d}{D} \cos \phi \right\}, \tag{1}$$

where f_r is the shaft speed (frequency), n is the number of rolling elements, d is the diameter of the rolling elements, D is the pitch diameter , and ϕ is the contact angle. Other fault characteristic frequencies are the BPFO (ball pass frequency, outer race) and the BSF (ball spin frequency), corresponding respectively to outer race and rolling element faults.

In order to quantify the presence and severity of an inner race fault, we use as a feature the median of the SES value at the

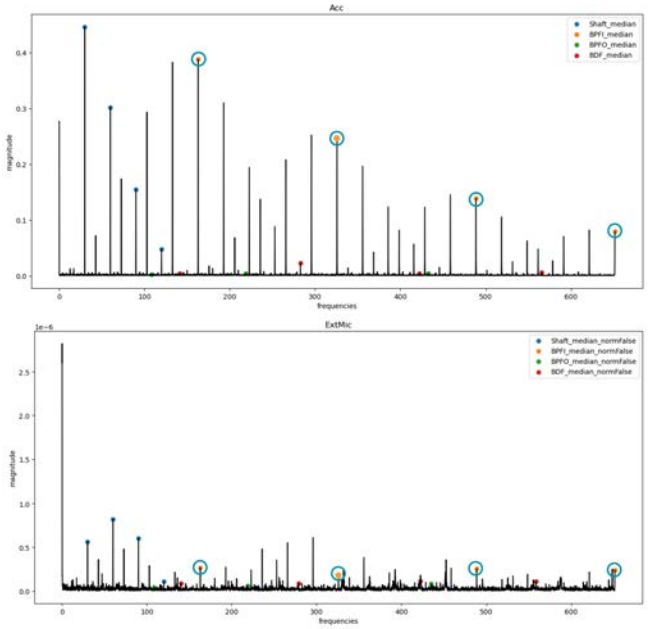


Figure 1. The squared envelope spectrum (SES) of the signal produced by a bearing with an inner race fault, acquired with an accelerometer (top) and a microphone (bottom). The BPFI and its first three harmonics are encircled in blue. The shaft frequency, BPFO and BDF (twice the BSF) are indicated for the sake of completeness.

BPFI and its harmonics. Throughout the rest of the paper, we refer to this feature as the BPFI feature. Mathematically, it is expressed as

$$\xi_{\text{BPFI}} = \underset{k}{\text{median}} \left\{ Y(k f_{\text{BPFI}}) \right\}, \quad k \in \{1, \ldots, n_{\text{harm}}\}, \quad (2)$$

where $Y(k f_{\text{BPFI}})$ denotes the peak magnitude of the SES at the k-th harmonic of the BPFI, $\underset{k}{\text{median}}\{\cdot\}$ denotes the median value of the set indexed by the integer k, and n_{harm} is the number of harmonics considered. Finding the peaks is done by searching the maximum SES magnitude around the theoretical fault frequency (Eq. 1) and its harmonics, within a pre-defined range tolerance.

2.2. Noise-unaware acoustic diagnosis

The direct application of diagnosis based on the fault characteristic frequencies, such as the BPFI, to acoustic sensing presents two problems. The first is that microphone signals are generally weaker than vibration signals, due to the larger distance between the microphone and the bearing. An example of this issue is provided in Figure 1, which shows a comparison of the SES from an accelerometer and a microphone signal produced by a bearing with an inner race fault. The BPFI and its first three harmonics can be easily identified in the accelerometer SES, while they cannot be clearly distinguished from the noise floor in the microphone SES.

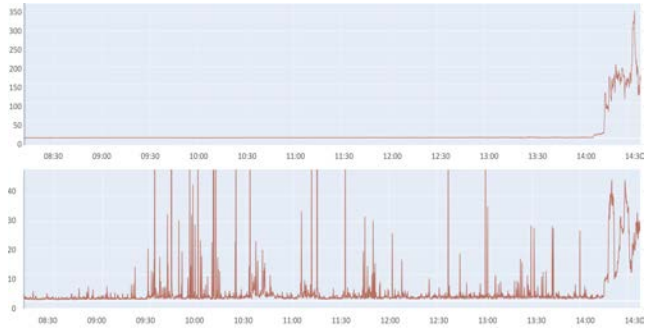


Figure 2. Comparison of the BPFI feature ξ_{BPFI} between an accelerometer (top) and a microphone signal (bottom) over a bearing accelerated lifetime.

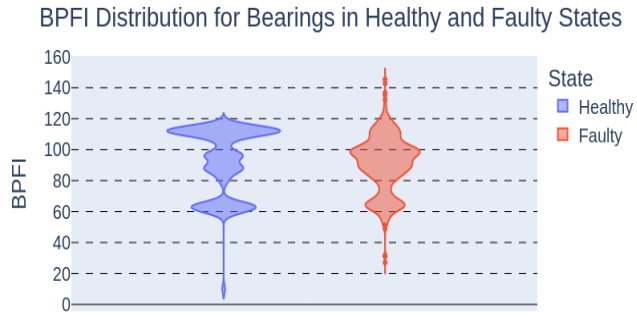


Figure 3. Comparison of the distribution of microphone-based BPFI feature ξ_{BPFI} between healthy and faulty states over the entire dataset.

The second problem is that the background acoustic noise is considerably stronger than the noise present in a vibration signal acquired by an accelerometer, and more diverse in nature due to the wide variety of potential noise sources present in industrial environments. As a result, microphone signals typically have a significantly poorer SNR. Moreover, due to this background noise, there will be additional energy present in the BPFI and its harmonics even when the bearing is healthy, leading to a great number of false positives over the bearing's lifetime.

This matter is illustrated in Figure 2, where a comparison is shown between the BPFI feature of an accelerometer and a microphone signal over the lifetime of a bearing in one of our accelerated lifetime experiments. The experimental setup and conditions are described in Section 4. It can be readily seen that the BPFI feature in the accelerometer signal displays a clear distinction between the healthy and faulty states of the bearing, while the BPFI feature in the microphone signal exhibits many spikes during the healthy state, leading to an unreliable diagnosis of the bearing inner race fault. Figure 3 further demonstrates the difficulty by showing the great overlap between the distributions of BPFI feature values acquired through the microphone for bearings in healthy and faulty states.

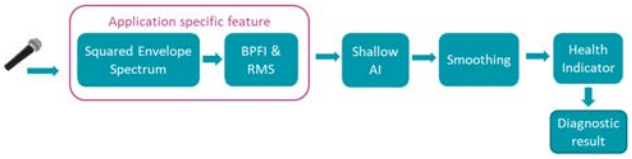


Figure 4. Diagram of the noise-unaware method with shallow AI for diagnosis of bearing inner race faults with acoustic sensing. The diagram would represent the baseline method by removing the shallow AI and the RMS feature.

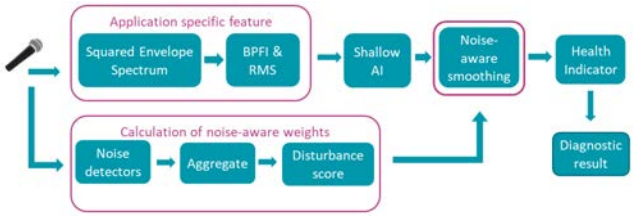


Figure 5. Diagram of noise-aware smoothing for diagnosis of bearing inner race faults with acoustic sensing.

The simplest strategy to enhance the diagnostic performance using the BPFI feature is to introduce a smoothing step. This method, with the choice of a median filter for smoothing, is what we consider our baseline for comparing the performance of diagnostic methods in this paper.

A more refined step is to introduce a shallow AI model before smoothing. In our case, this shallow AI is a two-layer fully connected neural network (NN) whose input features are the BPFI feature ξ_{BPFI} and the RMS value of the microphone signal. Both methods do not take the presence of acoustic noise explicitly into account, so we refer to them as noise-unaware methods. In particular, the BPFI and RMS features are very poor informants on the presence of background noise, hence the shallow AI can learn very little about rejecting undesired disturbances.

Figure 4 displays a diagram representing noise-unaware diagnosis with shallow AI and smoothing. The same diagram would represent the baseline method by removing the RMS feature and the shallow AI block.

3. Methods for robust acoustic bearing fault diagnosis

In this section we describe the two data-driven methods that we propose to increase robustness to noise in acoustic bearing fault diagnosis. As explained in Section 2, background noise introduces unreliability in the form of a high amount of false positives. The goal becomes therefore to reduce these false positives while retaining as much of the true positives as possible.

3.1. Noise-aware smoothing

Noise-aware smoothing aims to refine the health indicator calculated from the BPFI feature by taking into account the noise level present at each interval of time. To achieve this, a weighted median filter is applied to the raw health indicator over an interval of the last N points, where the weights are designed such that the influence of each point is inversely proportional to its noise level. A diagram of the diagnosis process including this strategy is shown in Figure 5.

The weighted median filter works as follows. If we are given a series of predictions $x_0, ..., x_t$ with noise levels $d_0, ..., d_t \in [0, 1]$, and a window size N, the noise-weighted prediction at time t would be computed as follows:

1. Assign a weight to every prediction $x_{t'}$ as $1 - d_{t'}$.
2. Sort predictions $x_{t-N}, ..., x_t$ and keep their associated weights $w_{t-N}, ..., w_t$ in that order too.
3. Compute the cumulative weight for each item $x_{t'}$, i.e. $\sum_{a=t-N}^{t'} w_a$.
4. The item where the cumulative weight exceeds half of the total weight is the weighted prediction, i.e. x'_t at time t such that $\sum_{a=t-N}^{t'} w_a \geq \frac{1}{2} \sum_{w=t-N}^{t} w_a$

Median smoothing follows the same procedure, except that all the weights are set to 1, reducing it to a regular median filter.

In order to obtain a noise level, each time interval is assigned a score that represents the likelihood that an undesired acoustic disturbance is present in it. For our case, an undesired disturbance is defined as any short-time sound that is not informative about the phenomenon being monitored, i.e., all sound events not generated by the bearing of interest itself. This excludes stationary background noise as well as disturbances that take last longer than a round of data acquisition (ten (10) seconds in our case).

This score is computed as the maximum of the output of a collection of noise detectors. These detectors are designed to indicate acoustic disturbances that can be characterized as events. This means any sounds whose presence in time, although it may be repeated, is limited. Examples include tools getting dropped, sporadic speech, various machinery turning on or off. Specifically, we implemented detectors targeting disturbances with the following characteristics:

Narrow-band disturbances: This detector indicates the presence of noise in a specific frequency band. In our experiments, the sources of this kind of disturbances were a pump and a forklift present in our laboratory.

High frequency disturbances: This detector indicates the presence loud, complex noises that have a lot of energy in high frequency bands. In our experiments, the source of this kind of disturbances was a neighboring experimental setup that kept loudly dropping off metal pipes.

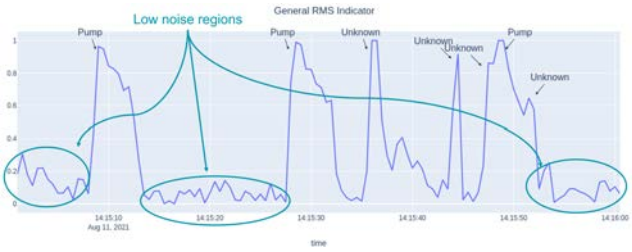


Figure 6. Example of the output of the detector for general loud disturbances. The regions circled correspond to intervals where no disturbance is detected.

General loud disturbances: A detector for loud events, designed to capture sudden changes in the root mean squared (RMS) value of the signal. This takes into account the fact that the bearings do not cause such changes at any point in their operation.

Speech disturbances: Voice Activity Detection (VAD) is a field of active research with many mature results. For this reason, we chose to utilize a VAD solution, provided by the Silero project (SileroTeam, 2021), based on a pre-trained neural network.

All detectors, except those for speech and general loud disturbances, require a characterization of the acoustic disturbances in the environment where the bearing of interest is located. An example of the output of the detector for general loud disturbances is shown in Figure 6. This detector captures some loud events that do not correspond to our known disturbance sources, marked as unknown. It also reacts to the pump activation, since it produces sudden changes in the RMS value of the signal. Capturing the same disturbance with several detectors is beneficial, as we are interested in catching as many as possible rather than determining their type. The regions of low disturbance score, that appear circled in the graph, are those given higher weight by noise-aware smoothing.

3.2. Noise awareness with deep AI and hybrid features

This method aims to utilize a deep AI model to obtain a reliable health indicator of bearing faults from acoustic information. The main idea is to introduce and train a deep AI model that uses adequately general features extracted from microphone recordings of healthy and faulty bearings. At a high level, it operates as a generalized way to clue the model in about what parts of the frequency spectrum are useful to pay attention to, and which parts are best to ignore. This model acts in combination with the very specific fault frequency features, thus integrating a data-driven technique with domain knowledge. The diagram in Figure 7 represents the diagnostic process that combines the indicators from both the BPFI feature and the deep AI-model.

The advantages of using a deep AI model are twofold. The first one is that it can learn complex patterns during the training process, leading to better diagnostic performance. The

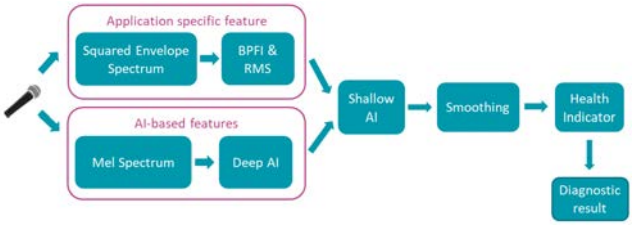


Figure 7. Diagram of deep AI-based noise awareness for diagnosis of bearing inner race faults with acoustic sensing.

second one is that, as long as sufficiently varied examples of disturbances are included during the training phase, it can learn to work with many kinds of noise sources.

The features for this method need to represent the relevant bearing fault information while being of reasonable dimensionality. For this purpose, the features we chose are the mel-spectrogram of the acoustic signal. This is a spectrogram obtained by a mel filter bank, a set of half-overlapped triangular filters equally spaced on the mel scale (Rabiner & Schafer, 2010). Since this is a logarithmic scale for frequency, the filters are narrower for low frequency bands and wider for high frequency bands.

For training, we chose a supervised approach, where we use as labels the output of anomaly detection from the accelerometer signal as ground truth. The features are normalized using their values at the start of the experiment, as the absence of normalization would be too sensitive to microphone gain and positioning.

There are several choices for the deep AI model, such as a deep neural network, a recurrent neural network, a temporal convolutional network or a transformer. In this work, our choice is a deep neural network (DNN), whose specifics are given in Section 4.3.

4. Experimental setup

The bearing datasets used in this study are collected in Flanders Make's Smart Maintenance Living Lab (Ooijevaar et al., 2019). This lab is developed as an open test and development platform and aims to support the adoption of condition monitoring technologies in the industry. It consists of seven identical drive train sub-systems. The setups are designed to perform accelerated lifetime testing of bearings and run bearings to their end-of-life. The accelerated lifetime test allows to create surface fatigue faults in bearings and monitor the fault evolution and accumulation during the (accelerated) life.

4.1. Bearing test rig and accelerated lifetime experiments

One of these experimental setups to perform the accelerated lifetime test is shown in the middle image of Figure 8. The setup comprises of a single shaft with a test bearing. The shaft is supported by a support bearing on each side. The test bear-

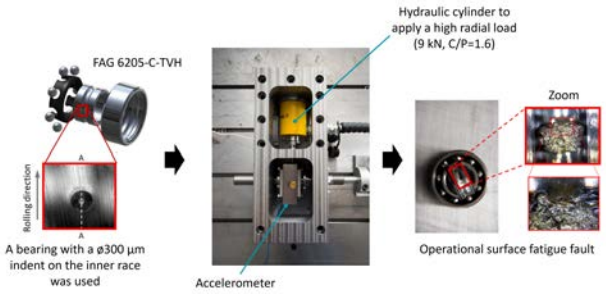


Figure 8. Illustration of the initial bearing state (left), the experimental test rig setup designed to perform accelerated life tests (middle), and the final state, a surface fatigue fault at the inner race of the bearing (right).

ing is lubricated by an internal oil bath. The setup is driven by a motor at a rotation speed up to 3000 RPM. In this work, we focus on experiments driven at 2000 RPM. Each setup is equipped with an accelerometer, temperature sensor, load sensor and speed sensor. The radial accelerations are measured at a sampling frequency of 50 kHz by an accelerometer attached to the bearing housing. The rotational speed and radial load of each setup can be controlled, such that each setup can operate at stationary and non-stationary operating conditions. An industrial Beckhoff control platform is used to acquire and store the sensor signals and to control the speed and load of each setup.

In total more than 70 bearing accelerated life tests have been performed on a FAG 6205-C-TVH deep groove ball bearing resulting in surface fatigue faults at the inner race. Two mechanisms are used to accelerate the bearing lifetime:

- A high radial load up to 9 kN (C/P = 1.6) is applied to the bearing outer ring with a hydraulic cylinder.
- Before the start of the test a small initial indentation of approximately 300 µm was created in the bearing inner race using a Rockwell C hardness tester. This indentation is used as a local stress riser and represents a local plastic deformation caused by, for instance, a contamination particle.

The accelerated life time tests are stopped as soon as 20g peak-to-peak accelerations are reached, resulting in severe rolling contact surface fatigue at the inner race (Halme & Andersson, 2009). The start and end condition of the inner race of one of the test bearing are shown in the left and right images of Figure 8.

4.2. Acoustic setup

The acoustic signals are acquired through two B&K 4189A21 microphones sampled at 50 kHz. One of them was placed under the safety cover of the bearing test setup, and the other outside of the cover, as showed in Figure 9. These microphones will be referred to respectively as *IntMic* and *ExtMic*

(a) Microphone inside the safety cover, referred to as *IntMic*.

(b) Microphone outside the safety cover, referred to as *ExtMic*.

Figure 9. Illustration of microphone positions used for the experimental recordings.

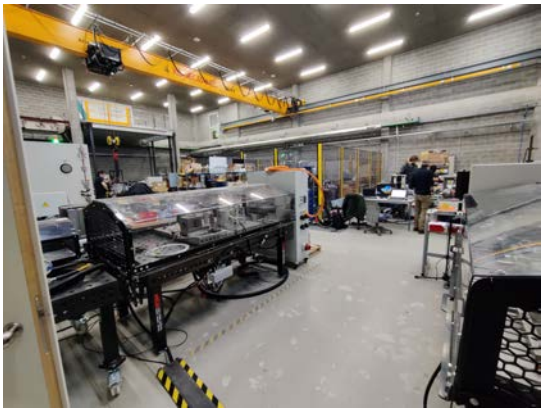


Figure 10. Illustration of the acoustic scene.

throughout the rest of the paper.

The experimental setup is situated in a large laboratory area at Flanders Make's facilities in Leuven, which has an uncontrolled and reverberant acoustic environment shown in Figure 10. It is a concrete room that contains many different kinds of setups such as drivetrains, looming machines etc., sometimes running simultaneously. There is also human activity with technicians and engineers running and maintaining the setups, or going about their daily activities. Due to the varying sizes of the setups here, sometimes small vehicles like forklifts or the crane integrated into the laboratory can also operate here. This makes the background noise potentially quite complex.

Specifically for the dataset we collected, there are a few common sources of noise that are often present, and we chose them as our focus for techniques that need us to characterize the kind of background noises that need suppressing. The most consistent, and arguably the simplest, disturbance is that coming from the hydraulic pump used to apply load on the test bearing. This pump activates roughly every minute in order to keep the pressure, and thus the load, constant. This cre-

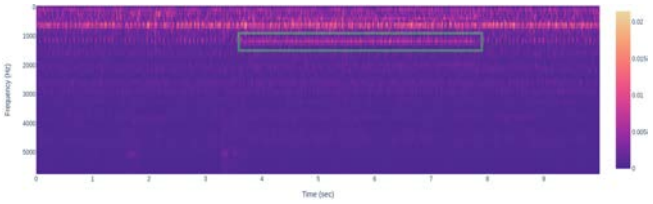


Figure 11. The spectrogram of an instance of pump activation with a faulty bearing being tested, captured by the *ExtMic* microphone. The pump activity is highlighted inside the green rectangle.

ates a very audible and consistent noise that is quite apparent in the signal, as can be readily observed in the spectrogram presented in Figure 11. The second common disturbance is a sharp, impulsive noise made by a nearby setup dropping metal pipes in a container. This happens once or twice a minute. Finally, on multiple occasions, there are people who are either walking or standing around the setup while in conversation. These are usually captured by the microphones, and constitute a third kind of disturbance we are interested in.

4.3. Processing configuration

The feature extraction is performed on segments of 10 seconds. For the BPFI feature, the tolerance of the SES peak search around the BPFI and its harmonics is set to 1.5% of the theoretical frequency (Eq. 1). The number of harmonics n_{harm} for the calculation of ξ_{BPFI} (Eq. 2) is set to 4. The mel spectrogram is calculated on 64 mel bands. The features are normalized using Z-score normalization, where the mean and standard deviation are computed on the first 30 minutes of the corresponding experiment. In experiments that lasted for more than one day, this calculation is done for each day.

Our dataset consists of two groups of run-till-failure experiments, all at 2000 RPM, where the bearing developed an inner race fault. The first group is characterized by continuous monitoring, where the sensors (accelerometer and microphones) constantly acquired data. This group contains 11 experiments, 7 of which run till failure and 4 of them ended prematurely. The second group contains 5 experiments were periodic monitoring of 1 second every 10 seconds was applied, 3 of them run till failure and 2 of them ended prematurely. This group is only used for training purposes. For cross-validation purposes, the dataset is split in three folds.

In addition to the captured data, a set of ground truth labels is also provided. It should be noted here that this labelling is not based directly on the physical state of the bearing, since it would not be available without stopping the test and dismantling the bearing, but based on analysis of the data captured by the accelerometer. Using this labelling the moment in time where the bearing starts having faulty behavior is determined. Data prior to this moment is then considered as healthy, and data afterwards is considered faulty.

For the deep AI model, we choose a deep neural network (DNN) with an input layer, three hidden layers of 32, 16 and 8 units, and an output layer.

5. Results and Discussion

In this section, we evaluate and compare the performance of both noise-aware and noise-unaware methods for bearing fault diagnosis using the microphone *ExtMic* signals, as its location outside of the safety cover of the setup is the most realistic. For clarity, we provide a summary of the methods evaluated in the following list.

1. **Noise-unaware methods**: These methods, described in Section 2.2, do not take the presence of noise explicitly into account. A diagram illustrating both methods is shown in Figure 4.
   (a) **Baseline**: The baseline method is based on the BPFI feature with median-smoothing.
   (b) **Shallow AI**: This method uses the BPFI and RMS features as input to a two-layer fully connected NN (shallow AI) and median-smoothing to achieve a diagnosis result.
2. **Noise-aware methods**: These methods aim to increase their robustness to noise, as explained in Section 3.
   (a) **Noise-aware smoothing**: The method described in Section 3.1, where the weights of the smoothing filter depend on the detected noise level. Its diagram is shown in Figure 5.
   (b) **Deep AI with hybrid features**: The method detailed in Section 3.2. It combines the BPFI and RMS features with mel spectrum features, where the latter are the input of the DNN (the deep AI model). It uses median-smoothing to obtain a diagnosis result. Its diagram is shown in Figure 7.
   (c) **Deep AI combined with noise smoothing**: This method is a combination of the two previous methods, i.e., the methods 2a and 2b.

5.1. Performance metrics

We use several metrics to assess the diagnostic performance of the proposed methods.

- **EPR**: The point on a precision vs. recall plot where these two metrics are equal. A high score indicates a high ratio of true positives with respect to both predicted positive samples and real positive samples. Expressed as a percentage.
- **ROD**: Rate of detection, the rate at which faults are detected before the safety stop, at a given precision value, equivalent to *recall*. In our case, we compute this value at 99% precision. More formally, if TP and FN denote respectively the true positive and false negative counts,

then

$$ROD = \frac{TP}{TP + FN}, \tag{3}$$

where $\frac{TP}{TP+FP} > .99$. Expressed as a percentage.

- **TOFD**: Average time of first detection in seconds, i.e, the time between the occurrence of the fault and the first time the model detects its, for a given precision value. In our case we compute this value at 99% precision, and it is only considered if the ROD is 100%. In the same way as our ground truth labels, the occurrence of the fault is defined based on accelerometer data. Note that this metric can be defined w.r.t. any source more reliable than the acoustic signal we are using. Formally, this metric is computed as

$$\frac{1}{N}\sum_{n=1}^{N} t_{fault,n} - t_{detection,n}, \tag{4}$$

where $ROD = 1.0$, $t_{detection,n}$ is the time of detection, and $t_{fault,n}$ is the time the fault occurred, both for the n-th experiment.

Regarding the TOFD metric, note that there is always some delay between the occurrence of a fault and its detection. There are two main sources of this delay. The first one is related to smoothing, and is not affected by the fact that we are running accelerated lifetime tests. This means that it would remain constant (for a given smoothing filter) even in regular testing. The second is the time gap between the fault being available to a vibration sensor versus it being available to an acoustic sensor. This delay pertains to the evolution of the fault, and therefore scales with the fault accelerations applied during the testing procedure. If we were to run regular lifetime tests, these delays would be multiplied by a corresponding factor. This means that, while the first delay is adjustable, the second delay is a consequence of the physics of the system and can only be reduced so much. It is a hard constraint on acoustic monitoring.

5.2. Performance results

The performance metrics for our evaluated methods are displayed in Table 1.

5.2.1. Performance of noise-unaware methods

It can be clearly seen that the baseline method's performance is quite poor, since its EPR is barely over 50 %, and at 99% precision it is only able to detect 14% of the faults.

The introduction of the shallow AI causes a significant jump in model performance, where it can now reach 100% ROD at a precision level of 99%, and an EPR point of 69%. However, note that the TOFD of 1520 seconds, about 25 minutes, is quite high, due to the required size of the smoothing window to reach a high precision. An example from the health indicator obtained by the shallow AI method is displayed in Figure 12, both before and after median smoothing. It can be observed that the shallow AI allows for a clear distinction of the faulty state when the fault actually develops, but it is the smoothing that removes the high amount of false positives. However, each process introduces a noticeable delay in the health indicator, which is expected to be dependent on the particular characteristics of the background noise.

Table 1. Performance results of the evaluated methods.

Method	EPR	ROD	TOFD
Baseline	55%	14%	-
Shallow AI	69%	100 %	1520 sec
Noise-aware smoothing	72%	100 %	1031 sec
Deep AI & hybrid features	79%	100 %	605 sec
Deep AI & hybrid features + noise-aware smoothing	81%	100 %	600 sec

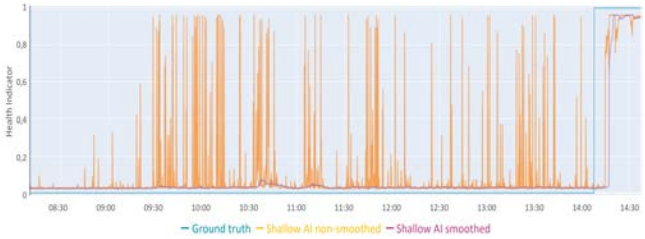


Figure 12. Example of health indicator obtained from the shallow AI method before and after median smoothing.

5.2.2. Performance of noise-aware methods

The use of noise-aware smoothing increases the EPR to 72%, maintains the ROD of 100%, and its TOFD is decreased by 33% with respect to the TOFD of the shallow AI method. This is a strict but moderate improvement over the best noise-unaware method.

The further addition of the mel spectrum features and the deep AI model causes a significant leap in performance, where the EPR point reaches 79%, and the TOFD is decreased by 42% and 60% of the TOFD values of the noise smoothing and shallow AI methods respectively. This improvement demonstrates the ability of the DNN to learn complex patterns from the mel spectrum features, and to complement the BPFI feature to achieve a better diagnostic performance. An example from the health indicator obtained by this method is displayed in Figure 13, both before and after median smoothing. It can be readily observed that before smoothing, the false positives are notably less frequent than in the example of the shallow AI method from the same experiment, shown in Figure 12. Smoothing removes these false positives, but crucially it in-

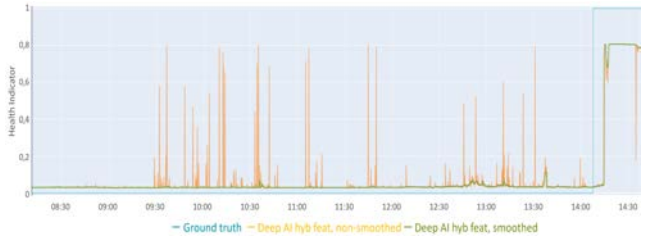


Figure 13. Example of health indicator obtained from the deep AI with hybrid features method before and after median smoothing.

Table 2. EPR points for median and noise-aware smoothing for methods 1b, 2a, 2b and 2c. Hybrid refers to the combination of BPFI and RMS with mel spectrum features.

Smoothing type	Features	
	BPFI & RMS	**Hybrid**
Median	69%	79%
Noise-aware	72%	81%

troduces less delay than in the shallow AI example. The reason is that, as the deep AI method produces fewer false positives, the smoothing window can be shorter, thus minimizing the additional delay introduced by this step.

Last of all, the combination of the two noise-aware methods, i.e., where noise-smoothing is applied to the deep AI method with hybrid features, achieves a moderate increase of the EPR until 81%, and no significant reduction in TOFD.

5.3. Effect of smoothing: noise-aware vs median

Note that the inclusion of noise-aware smoothing results in a moderate EPR improvement over the same method using median smoothing. This can be seen in Table 2, where the EPR points are arranged depending on the smoothing type and features that methods utilize. The reason is that noise-aware smoothing addresses the following issues:

- For a true increase in the anomaly score, noise-aware smoothing is typically faster to respond due to its non-even weighting.
- In case of quickly fluctuating anomaly scores, median smoothing is a lot less stable due to its inability to choose what to prioritize.

In most other cases noise-aware smoothing behaves comparably to median-smoothing, offering the same benefits. This makes noise-aware smoothing an attractive enhancement, although it comes with the additional cost of designing appropriate noise detectors for the acoustic scene where the machinery of the monitored bearings operates.

6. Conclusion

In this study, we have focused on developing robust methods for acoustic condition monitoring of inner race faults in rolling element bearings in industrial environments. This is a challenging problem due to the strong influence of background noise, which introduces a high amount of false positives and delays fault detection, resulting in poor diagnostic performance. Our two proposed noise-aware methods have different levels of complexity. The first and simplest one is noise-aware smoothing, which adapts the smoothing weights according to the detected noise levels. The second and more complex one is a deep AI model that uses mel spectrum features and acts in combination with the bearing fault frequencies to achieve a diagnostic result. These methods have been trained and tested with an accelerated bearing lifetime dataset acquired in the Flanders Make Smart Maintenance Lab, which is a reverberant environment where strong and diverse acoustic disturbances were present.

The results demonstrate significant improvements over the noise-unaware baseline, both in diagnostic performance and in detection time, using a single microphone signal. Moreover, these benefits are distinct both when the noise-aware methods are applied independently or in combination, so they can thus be chosen according to the monitoring requirements of each particular use case. In summary, we have shown that, when employing adequate strategies to increase robustness to noise, acoustic monitoring can be a cost-effective and practical alternative for vibration monitoring. Future work in this problem involves studying the influence of the training dataset size on accuracy, applying and testing our strategies to other bearing fault types, and studying the effect of data augmentation in the training of the deep AI model.

Acknowledgment

This research work was supported by Flanders Make, the strategic research centre for the manufacturing industry, and more precisely by the ACMON ICON research project. Authors would like to thank this project for funding the research presented in this paper.

References

Cardenas Cabada, E., Leclere, Q., Antoni, J., & Hamzaoui, N. (2017). Fault detection in rotating machines with beamforming: Spatial visualization of diagnosis features. *Mechanical Systems and Signal Processing*, *97*, 33-43. (Special Issue on Surveillance)

Glowacz, A. (2019). Fault diagnosis of single-phase induction motor based on acoustic signals. *Mechanical Systems and Signal Processing*, *117*, 65-80.

Halme, J., & Andersson, P. (2009). Rolling contact fatigue and wear fundamentals for rolling bearing diagnostics -

state of the art. *Journal of Engineering Tribology*, *224*, 377–393.

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mechanical Systems and Signal Processing*, *42*(1), 314-334.

Mian, T., Choudhary, A., & Fatima, S. (2022). An efficient diagnosis approach for bearing faults using sound quality metrics. *Applied Acoustics*, *195*, 108839.

Ompusunggu, A. P., Devos, S., & Petre, F. (2013). Stochastic-resonance based fault diagnosis for rolling element bearings subjected to low rotational speed. *International Journal of Prognostics and Health Management (IJPHM)*, *4*.

Ooijevaar, T., Pichler, K., Di, Y., Devos, S., Volckaert, B., Hoecke, S. V., & Hesch, C. (2019). Smart machine maintenance enabled by a condition monitoring living lab. *IFAC-PapersOnLine*, *52*(15), 376-381. (8th IFAC Symposium on Mechatronic Systems MECHATRONICS 2019)

Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing* (1st ed.). USA: Prentice Hall Press.

Randall, R. B. (2011). *Vibration-based condition monitoring: Industrial, aerospace and automotive applications*. John Wiley & Sons, Ltd.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—a tutorial. *Mechanical Systems and Signal Processing*, *25*(2), 485-520.

Ricardo Mauricio, A. M., Denayer, H., & Gryllias, K. (2022). Time-domain beamformed envelope spectrum of acoustic signals for bearing diagnostics. In *Conference proceedings of ISMA 2022 - USD 2022.*

Ricardo Mauricio, A. M., Denayer, H., & Gryllias, K. (2023). Beamformed envelope spectrum of acoustic signals for bearing diagnostics under varying speed conditions. In *Proceedings of NOVEM 2023.*

SileroTeam. (2021). *Silero models: pre-trained enterprise-grade STT / TTS models and benchmarks.* `https://github.com/snakers4/silero-models`. GitHub.

On the Feasibility of Condition Monitoring of Belt Splices in Belt Conveyor Systems Using IoT Devices*

Henrik Lindström[1], Johan Öhman[2], Vanessa Meulenberg[3], Reiner Gnauert[4], Claus Weimann[5], and Wolfgang Birk[6]

[1,2,3,6] Predge AB, Västra Varvsgatan 11, 97234 Luleå, Sweden
{Henrik.Lindstrom,Johan.Ohman ,Vanessa.Meulenberg, Wolfgang.Birk}@predge.se

[6] Automatic Control, Luleå University of Technology, 97187 Luleå, Sweden
Wolfgang.Birk@ltu.se

[4,5] HOSCH Fördertechnik GmbH, Am Stadion 36, 45659 Recklinghausen, Germany
{Reiner.Gnauert, Claus.Weimann}@hosch.de

ABSTRACT

This paper investigates fully automated condition monitoring of belt splices within operational belt conveyor systems, using IoT devices to predict and inform on potential belt breakage or tearing. Such events cause production stops and potentially harm workers. Belt splices are laminated belt connections subject to deterioration during operation and are usually weak spots. The proposed scheme circumvents manual inspection efforts and uses the HOSCH^{iris} DISCOVER IoT device for sensing and data acquisition. Each belt conveyor is equipped with one individual IoT device acquiring the motion signal of the scraper which is used to learn signal patterns of the pulley and the belt to identify both location and deterioration of the individual splices. Deterioration is characterized from an initial healthy condition to a severe condition of the splice to inform on the potential need for action. To assess the feasibility of the scheme, several tests are designed and performed in an industrial belt conveyor system. The results indicate that the scheme can provide valuable insights into the splice condition and its degradation.

1. INTRODUCTION

Belt conveyor systems are widely used to transport material and are an essential component in many industry sectors but are often critical assets in a production chain of bulk material, like in e.g. mining. Unexpected breakdowns of belt conveyor systems render production stops, losses, and can severely harm workers in close perimeter of such events.

The belt usually consists of several pieces, vulcanized together to achieve sufficient length. The vulcanized joints are called splices. In Figure 1, a simplified sketch of a belt conveyor is shown, and how splices could be distributed along the belt. The condition of these splices deteriorates during operation, leading to breakage or tearing. To preventively detect damage, all splices are regularly inspected. For this, the belt is run empty and at low speed to visually assess the belt surface and splices by a worker, leading to frequent downtimes in production. The quality of this manual condition assessment depends on the workers' expertise to identify issues on a moving belt, while keeping sufficiently attentive and tracking the splice locations. Such a campaign can last for several hours for longer belt conveyors, and thus human errors are not uncommon. To circumvent the problem of production losses, there is a need for monitoring solutions that work during normal operation.

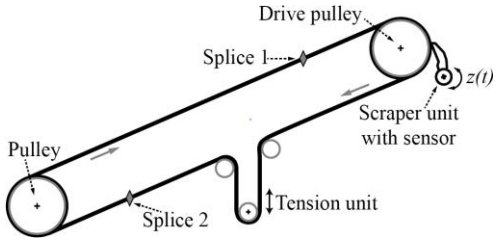


Figure 1: Sketch of the belt conveyor system with two splices and a scraper unit equipped with the sensing device generating the output $\boldsymbol{z(t)}$.

Various methods have been proposed for monitoring steel cord belt splices. Min (2010) suggests using Hall effect sensors to measure belt deformation and bending moment equations to assess the tensile force. However, concrete results validating this technique are lacking. Harrison (1985) and Kozłowski et al. (2020) propose methods based on measuring magnetic fields generated by belt reinforcement steel cords. Kozłowski, et.al. (2020) found that through a

variability analysis, the magnetic current can be compared to an estimated pattern. With this method, the cords could be monitored. Bancroft et al. (2017) used a camera and encoder to visually inspect mechanical splices and could determine the splice condition. Alport et al. (2001) developed artificial neural networks for splice monitoring using conveyor belt video footage, achieving splice identification accuracy of at least 89%. However, further development is needed for monitoring splice degradation and internal defects. Roxon Oy's HX products utilize laser scanning for surface damage and for monitoring the belt thickness that could detect splice elongation (Roxon, 2024).

The solutions discussed above require splices with steel cord reinforcement or mechanical splices that are easily visible. A camera installation or laser scanner, even though it has potential, requires the additional installation of equipment, resulting in increased maintenance costs. Moreover, the methods described above frequently lack definitive results regarding their accuracy.

The contribution of this paper is an analytics solution for condition monitoring of belt splices utilizing the displacement data from one individual belt scraper, as patented by Weimann and Kiel (2020). The benefit of this approach is that it can be used for all belt configurations, while in operation and with material on the belt. Since only the displacement of the scraper is analyzed, no additional equipment is required, making it a cost-effective solution.

The paper is structured as follows. First, a problem definition and description of the approach is given, followed by a summary of the scraper sensing solution. Next, the analytics solution is described, including belt speed estimation, transformation to distance domain, belt signature estimation, splice re-identification and degradation, and condition estimation. Thereafter, the proposed method is applied and tested in a real-life setting and the results are presented and discussed. Finally, the work is concluded.

2. Problem Definition and Approach

Splices are fixed locations on a belt which means that the individual splice locations need to be re-identified in $z(t)$, which is the displacement of the scraper (Figure 1). While $z(t)$ is time-based, the splice itself has a spatial location and structure along the belt. The problem of the condition monitoring of a splice over time is therefore to identify the passage of an individual splice at the sensor and to assess its degradation based on the signal that is acquired during the passage of the splice at the sensor. Moreover, the belt speed is not constant and needs to be treated as unknown. Using an individual IoT device combined with the scraper to measure $z(t)$ would avoid any integration of the sensing solution with the control system or IT infrastructure, making deployment easy and fast.

The approach to address the problem is as follows: The HOSCH^{iris} DISCOVER System is selected as the IoT device measuring the displacement $z(t)$ of the scraper. From the measurement and using design information of the pulley, the belt speed is estimated. Thereafter, the measurement signal $z(t)$ is transformed into the spatial or distance domain, denoted $z(d)$, where d denotes the distance that is covered. The annotated locations of the splices in the distance domain can then be re-identified in $z(d)$ requiring the detection of complete belt revolutions in the data. For every revolution of the belt, the splice locations can then be assessed, and their change can be tracked over the number of belt revolutions or time. Using the change in $z(d_i)$ for splice i at location d_i, condition indicators are derived and then mapped into actionable insights for decision making on maintenance or stopping of the belt conveyor.

Some challenges to this approach must be addressed. First, the measurement signal is affected by noise, which come from the surface structure of the belt and the pulley, but also from the scraping action to remove material from the belt. It is also not uncommon that material can get stuck between the scraper and the belt for some time which can lead to a temporary large displacement signal. How these effects will be managed is described in Section 4.

Moreover, the solution is intended to work independently of a control system or any integration into the IT infrastructure of the belt conveyor owners. The solution is therefore implemented in a cloud-based architecture as depicted in Figure 2. There, the IoT device connects with the HOSCH cloud to ingest the data and makes it accessible for the analytics in the partnering Predge cloud. All front-end functionalities are collected in HOSCH cloud, like configuration, alarming, visualization of actionable insights on the splices, and dashboards for the decision making of the user.

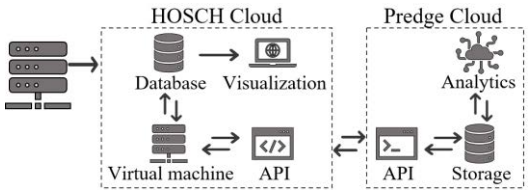


Figure 2: Cloud architecture to acquire the IoT device data, store and process it to provide actionable insights.

3. Sensing Solution

In this section, a short description is given of the sensing solution, based on the patent by Weimann et al. (2020). Figure 3 depicts a sketch of the scraper at the drive pulley on the belt, including the sensor. The pulley has a lining generating a high friction surface that is in contact with the belt. In the current setup, the high friction surface consists of three segments. The scraper is fixed to an axle where it can rotate. Connected to the scraper is a spring rod to adjust the

tension at which the scraper is in contact with the belt. The spring rod will move proportionally to the scraper. Opposite the spring rod, there is a sensor mounted to the housing which measures the distance between the sensor and a magnet attached to the end of the spring rod.

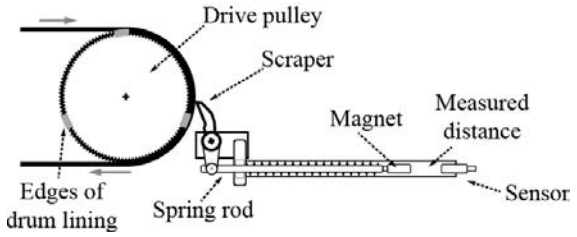


Figure 3. HOSCH HD PU pre scraper connected to the pre-tensioning spring and sensing device.

In **Error! Reference source not found.**, the raw sensor s ignal is displayed showing approximately 16 meters of the belt passing by the sensor, where the speed was increased at 12:22:07. The recurrent sinusoidal-like motion originates from the surface structure of the pulley where the drum is equipped with a lining composed of several segments. It is important to note that the pulley surface is a disturbance in the signal and may mask the belt surface changes. When the speed is increased the pulley rotational speed is increasing which leads to an increase in the frequency of the sinusoidal-like disturbance.

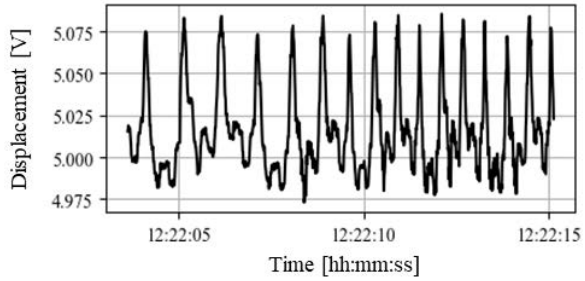


Figure 4. Raw sensor signal for a short time where the conveyor belt is run at two different speeds.

The sensor is connected to the IoT measurement system that samples the sensor signal and pre-processes it. It is thereafter locally stored and transmitted to the cloud wirelessly.

4. ANALYTICS SOLUTION

In this section the analytics solution to determine the condition of splices on a belt is presented.

4.1. Overview

The condition monitoring of splices using the scraper displacement measurement $z(t)$ requires several steps, as shown in Figure 5. One reason for this is that the splices are not as prominent in the sensor signal as the joined effect of all disturbances, like sensor noise, surface structures of pulley and belt, and displacement due to material removal from the belt. Consequently, the algorithm needs to recover the displacement that is attributed to the splices. In addition, the displacement needs to be correctly assigned to an individual splice to assess the surface change at the splice location.

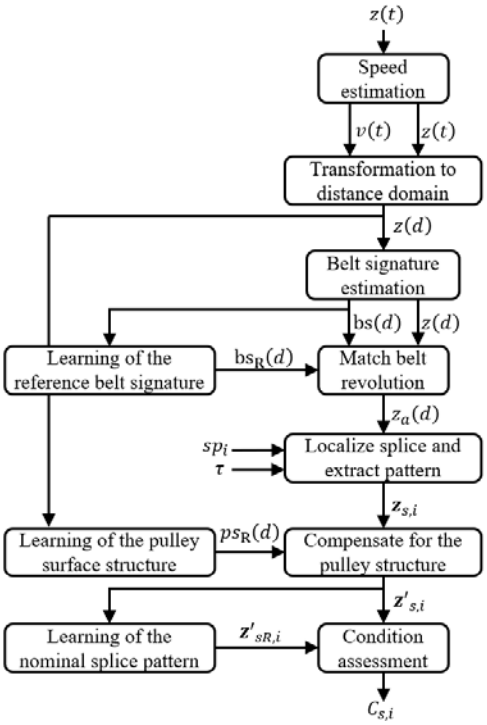


Figure 5. Block diagram for the condition monitoring of splices.

To ensure that surface changes are correctly assigned, the measured displacements need to be associated with specific locations along the belt. By estimating the belt speed from the measurement signal, transforming it into the distance domain and then identifying complete revolutions of the belt, it is possible to associate displacement measurements with specific coordinates along the belt. The identification of complete revolutions is done using an estimate for the belt surface signature, denoted $bs(d)$.

After an initial learning of a reference belt signature $bs_R(d)$, it is possible to match a complete revolution. Since splices are fixed locations sp_i along the belt with an overestimated length τ, a distance series can be extracted from the aligned raw data $z_a(d)$, reflecting the splice location, denoted $\boldsymbol{z}_{s,i}$. Since the pulley rotation and belt rotations are not aligned, the displacement induced by the pulley surface structure needs to be compensated for rendering $\boldsymbol{z}'_{s,i}$. Now the change in the surface can either be assessed in absolute terms or relative to a nominal $\boldsymbol{z}'_{sR,i}$, which is learned from data or provided by the user. The surface change can be assessed in different way and renders a condition indicator $C_{s,i}$. An advantage of this approach is that local belt damages are not confusing the association of splices to specific data series which enables the detection of new monitoring locations.

4.2. Speed estimation

To analyse the splice condition, the belt speed $v(t)$, needs to be estimated. The edges of the lining segments on the pulley introduce a distorted sinusoidal structure to the displacement data $z(t)$. This reoccurring structure in the signal is utilized when estimating the speed of the belt, by measuring the time between two registered edges and converting the time to a velocity. The displacement data $z(t)$ is split into five second batches of data $X(t)$, such that small changes in speed can be registered. $X(t)$ is now assumed to be a stationary process and can be normalized as

$$X' = \frac{X - \bar{X}}{\sigma} \tag{1}$$

where $\bar{X}$ is the mean and σ is the standard deviation of X. The normalization in (1) is done such that when computing the autocorrelation function (ACF) of the batch, there will not be any variance offset present. The ACF for a stationary process is defined as

$$r_{X'}(\tau) = \int_{-\infty}^{\infty} X'(t)X'(t+\tau)\mathrm{dt}. \tag{2}$$

The ACF in (2) is used here because of the repetitiveness of $z(t)$. The ACF will show local maxima at distances away from zero corresponding to the time it takes for the next lining segment to appear. Let i' be the solution to the following minimization problem that searches for the index in $r_{X'}(\tau)$ that corresponds to the first local maxima.

$$\begin{aligned} \text{Minimize} \quad & i \in \mathrm{dom}(r_{X'}) \\ \text{s. t.} \quad & i \geq m \\ & r_{X'}(i) \geq h \\ & \nabla r_{X'}(i) = 0 \\ & \Delta r_{X'}(i) < 0 \end{aligned} \tag{3}$$

In (3), m is a lower limit of i, h is a lower limit of $r_{X'}$ at index i and the third and fourth criterions requires i to be at a local maximum to $r_{X'}$. The solution i' to (3) is then the smallest index which satisfies all the criterions for (3). The velocity of the belt is calculated as

$$v = \frac{D}{3i'} \tag{4}$$

where D is the circumference of the pulley. The calculations in (2), (3) and (4) are done for each batch, resulting in a vector of speed estimations that will later be used for transforming the time series into a distance series.

4.3. Transformation to distance domain

If the average speed in the speed vector was greater than 0.2 m/s, $z(t)$ is transformed into a distance series, $z(d)$. The time series signal is sampled at a fixed rate at instance k independent of the belt speed. The covered distance d_k by the belt is a multiplication of the time instances t_k by the belt speed $v(t_k)$. The resulting distance dependent series $z(d_k)$ is not sampled equidistantly. By applying a linear interpolation with a fixed distance sample rate of 1 cm, an equidistant distance series is found. The benefit of transforming the time series into a distance series is that it enables the comparison of splice data regardless of the belt speed, since the position of the splices and the pulley will always be the same.

4.4. Estimating the Belt Signature

Now that the measured signal is available as a distance series, it is possible to relate a specific position on the belt with a specific point in the distance series, if the starting point of the belt in the distance series is known. It is not necessary to know an exact starting point, but it should be known where a revolution of the belt starts and ends. The belt itself has a surface structure that will produce displacements at the scraper. This displacement will occur repeatedly in the distance series. However, the distance series is affected by disturbances, like e.g. the pulley surface structure, damage to the belt and operation related disturbances. Understanding the stochastic nature of the disturbances, the sinusoidal disturbance behavior of the pulley, and assuming the belt surface is smooth a Kalman filter can be employed to estimate the belt surface and its derivative $dbs(d)$. Note that the Kalman filter is not realized in the time domain but in the distance domain.

The underlying model for the Kalman filter is defined as a sinusoidal motion which is biased by the surface signature

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 1 & d_S & 1 & 0 \\ -\omega^2 d_S & 1 & 0 & 0 \\ 0 & 0 & 1 & d_S \\ 0 & 0 & 0 & 1 \end{bmatrix} x_k + \nu_k \\ z_k &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} x_k + \eta_k \end{aligned} \tag{5}$$

where ω is the spatial frequency of the sinusoidal-like motion induced by the pulley structure, d_S is the spatial sample rate, η and ν are normally distributed noise terms, and k denotes the sample instance. Further, the state vector is defined as

$$x = \begin{bmatrix} z & \frac{\partial z}{\partial d} & bs & \frac{\partial bs}{\partial d} \end{bmatrix}^T \tag{6}$$

The Kalman filter as described by Gustafsson (2000) is implemented using (5) as the model, initial conditions $x_0 = [z_0\ 0\ 0\ 0]^T$, and the variance of the sensor signal as R. Setting the covariance matrix Q reflecting ν and initial conditions for the state covariance matrix P, is usually difficult and dependent on the situation. Here, Q is chosen as a diagonal matrix $Q = diag(10^{-1}, 10^{-1}, 10^{-7}, 10^{-9})$ and the initial condition $P_0 = 100 \cdot Q$.

To identify belt rotations, the derivate $dbs(d)$ of $bs(d)$ is used in relation to a reference signature. Performing the estimation on several belt rotations enables the learning of a

reference signature by deriving the median of all recorded repetitions of the signature, denoted $dbs_R(d)$. Note, the reference derivative signature $dbs_R(d)$ can have an arbitrary starting point on the belt. To identify a complete revolution of the belt in the estimated signature, an optimization problem can be solved that identifies the position of $bs_R(d)$ in the currently estimated $dbs(d)$, by minimizing the deviation between the two series. The localization and identification of the splices and other points of interest (POI's) for monitoring is then solved as a lookup. The identification of the start of a belt revolution is describe here. Define

$$L' = \underset{L}{\operatorname{argmin}} \frac{1}{N}\sum_{i=0}^{N}(dbs(i+L) - dbs_R(i))^2 \quad (7)$$

where N is the number of datapoints in dbs_R. The solution L' to (7) will be an index where the reference dbs_R is the most alike dbs and will describe where a new belt rotation is taken place. By having a knowledge of the splice locations in belt reference $dbs_R(d)$, it is now possible to also localize the splice locations in $z(d_k)$.

4.5. Splice and Pulley References

Similar to having reference distance series for the belt signature, references for the splices and the pulley can be derived. Moreover, the pulley surface structure is a dominant disturbance of high magnitude and the rotation of the pulley, and the belt are only rarely aligned. Thus, one and the same position on the belt will be affected by different disturbances due to the pulley surface structure. Nevertheless, by having a pulley reference and aligning it with the recorded distance series, it is possible to remove it by subtraction from the distance series. As a result, the distance series representing the changes in the belt surface can be recovered. The remaining signal components are then $\boldsymbol{z'}_{s,i}$ and its reference $\boldsymbol{z'}_{sR,i}$. Using these two series it is possible to quantify the change in the belt surface and as a result calculate condition indices, that quantify the change over the number of revolutions of the belt.

4.6. Condition indices

The condition of the splice or any POIs on the belt can be characterized by two main parameters, the vertical displacement of the surface and the longitudinal extend of the area of change. Since the belt is composed of laminated layers of rubber and reinforcement materials, the lamination can degrade and variations in thickness can occur. Damages can also lead to lose parts or bubbles that can be filled with material. Typical condition indices include:

$$\left|\left|\boldsymbol{z'}_{s,i} - \boldsymbol{z'}_{sR,i}\right|\right|_2 \quad (8)$$

$$\max\left(|\boldsymbol{z'}_{s,i} - \boldsymbol{z'}_{sR,i}|\right) \quad (9)$$

$$\min\left(|\boldsymbol{z'}_{s,i} - \boldsymbol{z'}_{sR,i}|\right) \quad (10)$$

These indices can be tracked over time and their change can be predicted if the change is smooth over the number of revolutions of the belt. These condition indices can then be used as actionable insights for decision making on maintenance and stop of operation.

5. Results & Discussions

This section will present the results from tests that have been conducted at an industrial site. For this end, the solution was implemented in a cloud-based architecture as depicted in Figure 2. Two HOSCH[iris] DISCOVER units were installed on two belts with lengths of approximately 550 m and the splice condition monitoring scheme was adapted to the pulleys and belt. The tests were conducted to assess the speed estimation, belt signature estimation, and the condition monitoring.

For the condition monitoring specific tests were conducted, where rubber patches were glued to the belt surface. The splice areas themselves were newly vulcanized, which means they should not be estimated to a severe condition.

5.1. Speed Estimation

The estimated speed estimates were derived from the raw data signal by the algorithm. The displacement data is shown in Figure 6a where the belt starts from a stand still. The belt speed is then increased to 30%, 50%, 70%, 80%, 90%, 100%, and back to 30% of the maximum speed (3.3 m/s).

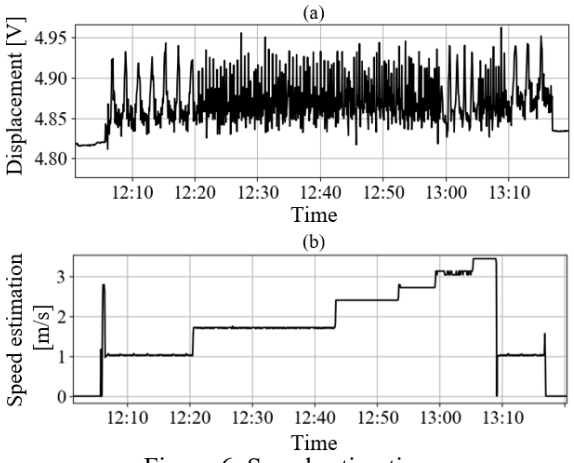


Figure 6. Speed estimation.

In Figure 6b, the estimated speed is shown. Shortly before 13:10, the speed is incorrectly estimated to be 0 m/s, which could be due to the loss of data in that time frame. There are also two spikes in the speed estimation at around 12:05 and 13:17, usually in high acceleration events when the belt is started or stopped. Using data from the control system, the speed estimation was validated rendering an error of 4% during the tests.

5.2. Belt Signature Estimation

For the belt signature estimation, no direct validation was possible, as the fine structure is estimated. Instead, the recurrence of the belt signature was used to assess the validity, by checking if the localization of the splices using the signature is correct.

In Figure 7a, a randomly picked sequence from normal operation is shown, where the displacement data is already transformed into the distance domain using the estimated speed. The estimated derivative belt signature $dbs(d)$, with its unique features can be seen in Figure 7b. The data shows two full revolutions of the belt and the noticeable similarity of the pattern before and after 60000 cm.

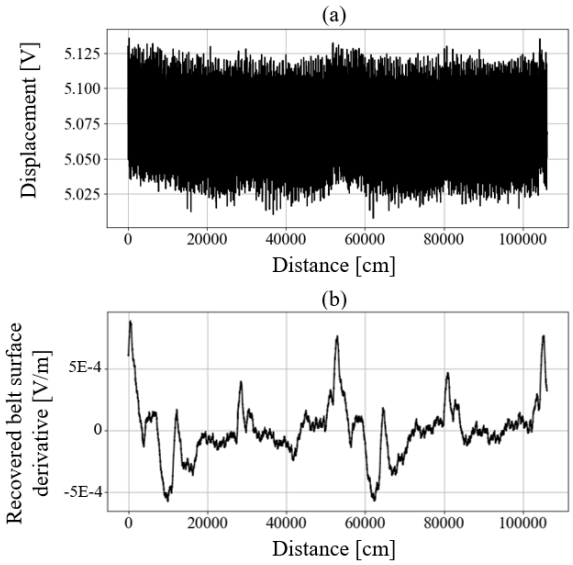


Figure 7. Transformation of displacement data to signature belt surface derivative

Using the pattern matching algorithm to align the belt rotations, the splices could be correctly localized within one meter of accuracy.

5.3. Condition monitoring of glued patches

To validate the condition assessment, artificially introduced POIs in the form of rubber patches glued to the belt surface were analyzed. For each belt, four patches of 100 x 200 x 1 mm were glued to the belt surface. The idea of the test was to track how the patches are degraded over the passages by the scraper and finally stripped from the belt surface.

To achieve this, references were created for each patch area and the already existing belt signatures and pulley references were utilized, which are generally true in the monitoring scheme. The belt conveyors were then operated as usual.

For the condition assessment of the patches the index in (9) is used and shown in Figure 8. The degradation of the patch is clearly distinguishable at about 13 belt rotations, and after about 21 rotations, it is no longer visible as it was removed from the belt by the scraper. This shows that POIs can be monitored and that changes in their behavior are estimated by in the monitoring scheme. However, the intensity of the patch degradation is not constant nor monotonically increasing each revolution and there is also some variation in the condition index.

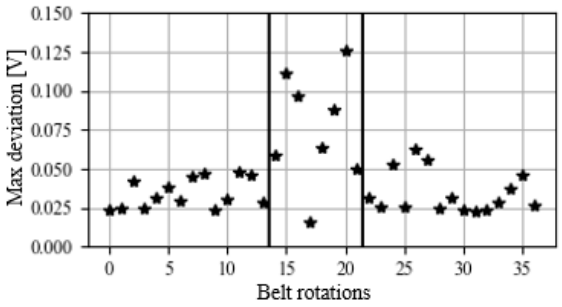


Figure 8. Maximum deviation of the belt before, while and after a patch was added.

Since the patches have sharp front edges, the collision induces an impulse on the scraper with subsequent movement. At the same time the scraper movement is sampled at a rate of 100 Hz. Depending on the alignment of the impulse with the sampling of the sensor signal and the belt speed, the maximum displacement might not be recorded, yielding a variability and error in the condition index. Nevertheless, the degradation phase of the patch is clearly captured by the scheme.

5.4. Condition monitoring of the splices

As already noted, the splices were newly vulcanized yielding a very smooth surface, which requires operation to take place over a long period of time (usually longer than a year). It was therefore expected that the splices would not generate any impact on the condition index. For the test, sequences from normal operation of the belt conveyor are used and information from inspections was collected.

Again, the maximum deviation index as given in (9) was used to assess the condition. The expectation from the test was that the index would not show any high values. In total 150 belt revolutions were assessed, which were received in batches of 10 minutes.

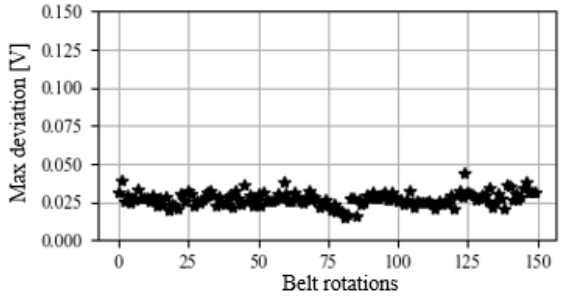


Figure 9. Maximum deviation of the belt for a splice area.

In Figure 9 it can be seen that no higher peaks are visible and that the condition index varies around a low value, which is

comparable to the lower value ranges that can be seen in Figure **8**. The inspections during the normal operation also confirmed that the splices are in good condition throughout the test period. It can therefore be concluded that the monitoring scheme is not generating false indications during normal operation and replicates the condition of the splices correctly.

6. CONCLUSION

In this work, the fully automated condition monitoring of belt splices within operational belt conveyor systems was investigated. It is shown how the belt speed can be estimated from a signal displacement signal, how a point of interest on a conveyor belt can be localized and its degradation can be monitored. For this end, typical statistical and Bayesian filtering approaches are applied together with simple learning schemes that provide data driven models of the belt, pulleys, and normal conditions of the points of interest.

Based on the conducted tests and their assessment it can also be concluded that the condition monitoring of the belt surface using the displacement signal of a single HOSCH^{iris} DISCOVER IoT device is feasible and that actionable insights on the degradation of the belt can be provided to operators and maintenance staff to ensure safe operation. The proposed solution is now online and part of the normal operation in an industrial plant.

Future work will target the collection of experience from the solution in normal operation, the benchmarking of used methods with other approaches, and assessing the effectiveness in capturing degradation events early on and in good time for decision making on operation and maintenance.

REFERENCES

Alport, M., Govinder, P., Plum, S., & Van Der Merwe, L. (2001). Identification of conveyor belt splices and damages using neural networks. *Bulk Solids Handling, 21*(6), 622-627.

Bancroft, B., Fromme, C., & Pilarski, T. (2003). Belt Vision System for Monitoring Mechanical Splices. *Proceedings of Longwall USA International Exhibition and Conference '03*.

Belt Condition Monitoring (2024). *Roxon.* Available at: https://roxon.com/hx-products/ (Accessed: 22 March 2024).

Gustafsson, F. (2000). Adaptive filtering and change detection (Vol. 1). New York: Wiley.

Harrison, A. (1985). A magnetic transducer for testing steel-cord deterioration in high-tensile strength conveyor belts. *NDT International, 18* (3), 133-138. https://doi.org/10.1016/0308-9126(85)90197-X

Kozłowski, T., Wodecki, J., Zimroz, R., Błażej, R., & Hardygóra, M. (2020). A diagnostics of conveyor belt splices. *Applied Sciences*, *10*(18), 6259. https://doi.org/https://doi.org/10.3390/app10186259

Min, H. (2010) 'Research on the Splice Breakage Monitoring System for Steel-Cord Belt Conveyor', *2010 International Conference on Measuring Technology and Mechatronics Automation*, pp. 223–226. doi:10.1109/icmtma.2010.86.

Weimann C, Kiel M (2020). Gurtabstreifer und Verfahren zum Betrieb eines Gurtabstreifers (German Patent No. DE102018123799A1)

BIOGRAPHIES

Henrik Lindström received his MSc. in Engineering Physics from Lund University, Sweden in 2022. He is currently working as an analytics developer at Predge AB. His work includes condition monitoring and health prediction of conveyor belts, railway turnouts and rail wagons.

Johan Öhman is an Analytics Developer at Predge AB. He holds an MSc in Engineering Physics (2015), Luleå University of Technology, and a PhD degree in Experimental Mechanics (2020), also from Luleå University of Technology. His research interests are physics-based, and data driven modeling of industrial systems.

Vanessa Meulenberg is an Analytics Developer and Project Manager at Predge AB. She completed her BSc. in Aeronautical Engineering from Inholland University of Applied Sciences and MSc. in Composite Materials at Luleå University of Technology.

Reiner Gnauert is the Head of Digital Business Development at HOSCH Fördertechnik GmbH. He holds an MSc. in Automation Control Avionics from Hamburg University of Technology. He has a long track record in conveyor belt monitoring and a history in validation and verification in the aviation industry.

Claus Weimann is the Head of Research and Development at HOSCH Fördertechnik GmbH. He holds an MSc. in Mechanical Engineering at RWTH Aachen University. He develops new products for conveyor belts at HOSCH with a focus on scraper technology.

Wolfgang Birk is the CTO at Predge AB and Professor of Automatic Control. He holds a M.Sc. degree in Electrical Engineering from University of Saarland (1997), a Ph.D. degree in Automatic Control from Luleå University of Technology (2002), and Professor of Automatic Control (2015). He has a background in the development of condition monitoring systems, process control systems for resource efficiency as well as active safety systems in the automotive sector. His research work has led to control and monitoring solutions in several industries. In the railway sector, his main interest and expertise is the use of on-board, way-side, and track monitoring systems for condition monitoring in operation and maintenance.

Particle Filter Approach for Prognostics Using Exact Static Parameter Estimation and Consistent Prediction

Kai Hencken[1], Arthur Serres[1,2], and Giacomo Garegnani[1]

[1] *Corporate Research Center, ABB Switzerland Ltd, Baden-Dättwil, CH-5405*
kai.hencken@ch.abb.com
giacomo.garegnani@ch.abb.com

[2] *current mail address:*
arthurserres1510@icloud.com

ABSTRACT

Particle filters are widely used in model-based prognostics. They estimate the future health state of an asset based on measurement data and an assumed degradation dynamics. Filters are in general applied to estimate only the states given a known dynamics of the process. In model-based prognostics, the dynamics is assumed to be known in an analytical form, but the parameters vary per device and need to be learned from the measurements as well. This is especially important for the calculation of the remaining useful life (RUL), as the prediction of the future evolution is needed.

There are commonly used approaches for this: Augmenting the state space with the parameter, together with assuming them to stay constant or adding an artificial diffusive evolution to them. The Liu–West filter improves on this by modifying the artificial evolution such that mean and standard deviation of the marginal parameter distribution are kept the same. Both approaches require to choose some tuning parameters, which might be difficult in practical applications. In addition, the model parameter is often assumed frozen for the prediction part, leading to an inconsistency. We propose how a modification of the parameter evolution in case of missing measurements can solve this in both cases.

More recently algorithms for combined state estimation and exact parameter estimation have been introduced, especially the Storvik filter, based on the usage of a sufficient statistic. We analyze how this can be applied to overcome difficulties with existing approaches, avoiding the need for tuning parameters. We also extend the Storvik filter in order to deal with time-steps with missing measurements. Two formally equivalent approaches are presented. These are applicable in all cases of missing measurements, coming either from irregular data acquisition, e.g. only during maintenance or inspection, or as part of the prediction step of the RUL calculation.

We study the different methods for two simple models in order to demonstrate potential issues with existing approaches and to explore the stability of the new one based on the Storvik filter. Finally we apply it to a practical application in the area of electrical distribution systems.

1. INTRODUCTION

The most common approach for predicting the end-of-life (EOL) of a device is to model its degradation. Let x_t be a degradation variable, describing the health of the device and evolving e.g. with time t. In the simplest case, we define the (soft) failure of the device as the condition that x_t reaches a predefined critical value x_{critical} (Goebel et al., 2017; Galar, Goebel, Sandborn, & Kumar, 2021). x_t can here be either a scalar or a vector, see e.g. (Peng, Ye, & Chen, 2018). We limit ourselves to the scalar case.

The evolution of x_t is in general described by a stochastic model. We restrict ourselves here to the discrete-time case, indexed by t, and assume that the state evolves from time instance t to $t+1$ as

$$x_{t+1} \sim p(x_{t+1} \mid x_t; \theta), \tag{1}$$

where p is a suitable probability model depending on a parameter vector θ. We assume the value of θ to be specific to each individual device rather than describing the behavior of a fleet and a prior distribution $p(\theta)$ to be known.

In most cases the degradation variable x_t is not directly measurable but needs to be inferred from an observable z_t. This might be a direct measurement of x_t corrupted by measurement error or a quantity that can be indirectly associated with it. Quite generally the relation between the degradation vari-

able x_t and the observable z_t is described by

$$z_t \sim p(z_t \mid x_t; \theta), \tag{2}$$

where p is a probability model possibly also depending on the parameter vector θ.

The combination of Eqs. (1) and (2) forms a state-space model (SSM) with unknown parameter θ.

The aim of prognostics is to calculate the remaining useful life given the measurements $z_{0:t_0}$ up to the current time horizon t_0, using the predictive probability

$$p(x_t \mid z_{0:t_0}), \quad t > t_0. \tag{3}$$

Note that in Eq. (3) the parameter θ does not appear, as it has been marginalized over. The uncertainty associated with θ is therefore automatically considered. Due to the stochastic nature of the evolution and the uncertainty in the parameter, the RUL is itself a random variable. Its distribution is given by

$$\begin{aligned} \mathrm{RUL}(t \mid t_0) \sim p(t \mid \{x_{t+t_0} \geq x_{\mathrm{critical}}\} \\ \cap \{x_{t'} < x_{\mathrm{critical}} \, \forall t' < t + t_0\}). \end{aligned} \tag{4}$$

A review of the prognostics paradigm and its applications can be found, e.g., in (Si, Wang, Hu, & Zhou, 2011; Jouin, Gouriveau, Hissel, Péra, & Zerhouni, 2016; Goebel et al., 2017; Galar et al., 2021).

Determining the RUL requires a combined state and parameter estimation approach. In principle, the estimation can be obtained with any Bayesian method, e.g., a Markov Chain Monte Carlo (MCMC) approach, that determines the joint distribution $p(x_{0:t}, \theta \mid z_{0:t_0})$ of all past, present and future states and the parameter using all measurements until time t_0. In practice, this approach is not viable, since the evaluation has to be repeated each time a new measurement point is added. Hence, the method becomes computationally more demanding as time increases. A sequential approach is more appropriate, such as a sequential Monte Carlo method (SMC), see e.g. (Doucet, de Freitas, & Gordon, 2001; Chopin & Papaspiliopoulos, 2020). This requires to update the joint distribution $p(x_t, \theta \mid z_{0:t_0+1})$ at each increase of the horizon taking into account only the new measurement z_{t_0+1} and the already known joint distribution until t_0. The determination of the distribution of the state $x_{0:t_0}$, especially of only the current state x_{t_0}, is a well-studied problem for known parameter values, regularly solved with the help of particle filters. In contrast, the determination of the joint distribution of states and parameter is a more difficult one and often solved by applying some approximations.

In addition to the estimation problem for x_{t_0}, prognostics applications require the ability to make predictions. Indeed, in order to compute the RUL as in Eq. (4) one needs to calculate the future distribution for x_t for $t > t_0$, see Eq. (3). A related topic is the ability to evolve the sequential approach when measurements are sparse and obtained at irregular time intervals only. For instance, we expect measurements to be of this form, if they are obtained as part of a maintenance or inspection routine. If the time interval between measurements is long, it is beneficial to evolve the distribution of state and parameter and only keep their values at the current time. It allows to continue to evolve the distribution up to the next measurement without the need to restart from the last measurement point.

In this paper, we consider the two issues above and present:

- A critical review of the state of the art on particle based sequential methods for joint state and parameter estimation in SSMs;
- Proposals on how to extend the methods to deal with missing measurements, required especially for future predictions;
- Explore the use of the Storvik filter as an exact approach to the combined state and parameter estimation problem for prognostics applications.

The remainder of this paper is organized as follows. In Sec. 2 we briefly review the application of particle filters to SSMs. In Sec. 3 we present common approaches to perform joint state and parameter estimation, introducing also a possible way to handle missing measurements. In Sec. 4 we review the Storvik filter, an exact method for the combined state and parameter estimation, together with an extension of the method in case of missing measurements in Sec. 5. In Sec. 6 we analyze the applicability of the methods for two simple models, that are typical for prognostics applications, and in Sec. 7 we present results of the application of the Storvik filter on real data. We conclude the paper with an outcome and give potential future directions in Sec. 8.

2. THE PARTICLE FILTER FOR SSMS

The particle filter is synonymous for the SMC approach to state estimation in nonlinear SSMs. For an introduction and review see, e.g., (Doucet, Godsill, & Andrieu, 2000; Doucet et al., 2001; Chopin & Papaspiliopoulos, 2020). Neglecting the parameter θ, particle filters approximate the probability distribution of the states by an empirical distribution based on a set of N particles $\{x_t^i\}_{i=1}^N$. The approximation incorporating weights for each particle is

$$p(x_t \mid z_{0:t}) \approx \widehat{p}(x_t \mid z_{0:t}) = \sum_{i=1}^{N} w_t^i \delta_{x_t^i}(x_t),$$

where the weights $\{w_t^i\}_{i=1}^N$ are normalized to sum to one, and where $\delta_x()$ denotes the Dirac delta distribution centered in x. Particles and weights are propagated and updated according to Bayes' rule using the SSM defined by Eqs. (1) and (2).

The simplest implementation is the bootstrap particle filter, given for completeness in Algo. 1. More complex algorithms have been proposed in the literature in order to overcome the shortcomings of the bootstrap particle filter in practice. The main is the impoverishment of the particle set. This refers to the fact that without resampling most weights degenerate over time, or with resampling (as done in the bootstrap particle filter) only a few particles are retained after the resampling step.

Algorithm 1 Bootstrap particle filter

1: Initialize $\{x_0^i, w_0^i = 1/N\}_{i=1}^N$;
2: **for** $t = 1, \dots, t_0$ **do**
3: **for** $i = 1, \dots, N$ **do** ▷ Propagate
4: Sample $\widetilde{x}_t^i \sim p(x_t \mid x_{t-1}^i)$;
5: Compute $\widetilde{w}_t^i = w_{t-1}^i p(z_t \mid \widetilde{x}_t^i)$;
6: **end for**
7: Normalize $\widehat{w}_t^i = \widetilde{w}_t^i / \sum_j \widetilde{w}_t^j$, $i = 1, \dots, N$;
8: **for** $i = 1, \dots, N$ **do** ▷ Resample
9: Sample k from the set $\{1, \dots, N\}$ with weights $\{\widehat{w}_t^j\}_{j=1}^N$;
10: Set $x_t^i = \widetilde{x}_t^k, w_t^i = 1/N$;
11: **end for**
12: **end for**

In this form, the particle filter assumes the model parameter θ to be known and fixed. Therefore, it cannot be directly applied to prognostics application, as the parameter values are typically specific to each individual device. Missing measurements can be treated on the other hand trivially: If the measurement z_t is unavailable, we marginalize Eq. (2) with respect to z_t and replace $p(z_t \mid \widetilde{x}_t^i)$ by $\int p(z_t' \mid \widetilde{x}_t^i) dz_t' = 1$. In practice, this corresponds to not updating the weights (line 5 of Algorithm 1) and only sampling $\widetilde{x}_t^i$ from the distribution $p(x_t \mid x_{t-1}^i)$. Resampling is not required in this case, but could still be done, even though it might lead to unnecessary impoverishment of the particles.

3. State and parameter estimation with particle filters

Estimating sequentially both state and parameter of a model is a difficult problem. Several methods have been proposed in the literature, see, e.g., (Doucet et al., 2000, 2001). Many of these do not treat the parameter estimation sequentially, and are therefore not further discussed here. In this section, we only describe three of the most common approaches in the prognostics literature, see, e.g., (Si et al., 2011; Jouin et al., 2016).

3.1. Parameter-augmented bootstrap particle filter

The straightforward approach to state and parameter estimation consists in augmenting the state-space x_t with the parameter θ, i.e., defining a new state space $X_t = (x_t, \theta_t)$. The underlying dynamics for x_t is unchanged and is defined by Eq. (3). The parameter θ is assumed to not evolve in time, i.e., it follows the trivial dynamics

$$\theta_t = \theta_{t-1},$$

and $\theta_0 = \theta$. Together with the prior distribution $p(\theta_0) = p(\theta)$ this is equivalent to the solution of the full problem. The augmented state of the resulting SSM can then be estimated using the bootstrap particle filter. The resulting method is given in Algo. 2.

With this method, the parameter θ_t does not evolve over time. Hence, the set of possible values for it is fixed throughout the algorithm, and is equal to the initial samples θ_0^i from the prior distribution $p(\theta)$. Due to resampling, only a few distinct values of θ survive after some time (in the worst case only one). Therefore, this algorithm leads in many cases to a strong overconfidence on the parameter uncertainty, and possibly to a wrong estimate of its value. Despite this shortcoming, this approach has been proposed in (An, Choi, & Kim, 2013), even if only in a tutorial setting.

Algorithm 2 Parameter-augmented bootstrap particle filter

1: Initialize $\{x_0^i, \theta_0^i, w_0^i = 1/N\}_{i=1}^N$;
2: **for** $t = 1, \dots, t_0$ **do** ▷ Propagate
3: **for** $i = 1, \dots, N$ **do**
4: Set $\widetilde{\theta}_t^i = \theta_{t-1}^i$
5: Sample $\widetilde{x}_t^i \sim p(x_t \mid x_{t-1}^i; \widetilde{\theta}_t^i)$;
6: Compute $\widetilde{w}_t^i = w_{t-1}^i p(z_t \mid \widetilde{x}_t^i; \widetilde{\theta}_t^i)$;
7: **end for**
8: Normalize $\widehat{w}_t^i = \widetilde{w}_t^i / \sum_j \widetilde{w}_t^j$, $i = 1, \dots, N$;
9: **for** $i = 1, \dots, N$ **do** ▷ Resample
10: Sample k from the set $\{1, \dots, N\}$ with weights $\{\widehat{w}_t^j\}_{j=1}^N$;
11: Set $x_t^i = \widetilde{x}_t^k, \theta_t^i = \widetilde{\theta}_t^k, w_t^i = 1/N$;
12: **end for**
13: **end for**

3.2. Diffusive bootstrap particle filter

The main limitation of the parameter-augmented bootstrap particle filter is the impossibility to create new parameter values θ_t^i. This can be overcome by increasing their variability over time and in particular by exploring values close to the particles that survive the resampling. Since we only have information regarding the likelihood function or posterior distribution of values of state and parameter represented by some particles, some approximation is needed.

The most popular approach to create variability in the parameter consists in adding a stochastic dynamic term to its time evolution. In almost all practical cases, this dynamics takes the form of a Brownian motion, i.e.,

$$\theta_t \sim \mathcal{N}(\theta_{t-1}, \Sigma_\theta), \tag{5}$$

where Σ_θ is a suitable covariance matrix.

Whereas the motivation for the stochastic evolution of the parameter θ is purely to improve the particle filter method, it is often proposed, that it attempts to capture the mismatch between the model and the real underlying process, even if the parameter is not – in principle – changing in time. Despite this mismatch being potentially a valid point, using a stochastic dynamics for this in prognostics is difficult to justify. Indeed, the variation of the degradation variable is often rather limited and deviations will tend to be rather systematic than random. Another case made is that it allows to capture change-points of the parameter, where the time evolution changes abruptly, e.g., due to a transition to a faulty state. Such transitions are often handled better using dedicated approaches. In addition, the diffusive nature of Eq. (5) leads to past measurements being considered progressively less by the filter, leading to a larger parameter and prediction uncertainty.

On a more practical side, introducing the covariance matrix Σ_θ adds hyperparameters to the algorithm that are often difficult to tune. Unfortunately, the performance of the algorithm relies strongly on a good choice of them. If the covariance has too small elements the parameter is essentially static and the method has the same issues as in Sec. 3.1. Conversely, if Σ_θ has too large elements, the dynamics introduces overdispersion to the parameter. This second case is particularly concerning in case of missing measurements and especially in the prediction phase. Indeed, without measurements, which are the driving force constraining the parameter, the diffusive dynamics leads to a strong and purely artificial increase in uncertainty. The calculation of the RUL is most susceptible to this, as a prediction over a long time horizon is made. A hybrid approach is often employed to overcome this difficulty: The parameter is evolved using the stochastic model for the estimation phase, but is then frozen for the prediction phase. This inconsistency is listed as one of the open questions in (Jouin et al., 2016).

We propose here the introduction of an improved parameter evolution by using a time dependent covariance matrix Σ_θ in order to mitigate this issue. The time dependence is defined in the following way: the parameter is only updated when measurements are done, otherwise it remains unchanged. This corresponds formally to setting $\Sigma_\theta = 0$ for time steps without measurements. With this practical approach, we do not incur an artificial but unneeded overdispersion and still retain the better exploration of the parameter space with respect to the method of Sec. 3.1. We also remove the inconsistency in calculating the evolution in the parameter estimation and prediction phase. Despite these improvements, the diffusive bootstrap filter still strongly relies on choosing appropriately the covariance matrix to avoid either the impoverishment of the particles, or the loss of information carried by past measurements.

The diffusive bootstrap particle filter algorithm including the improvement for missing measurements is shown in Algo. 3.

Algorithm 3 Diffusive augmented bootstrap particle filter including treating of missing measurements

Initialize $\{x_0^i, \theta_0^i, w_0^i = 1/N\}_{i=1}^N$;
for $t = 1, \ldots, t_0$ **do** ▷ Propagate
  for $i = 1, \ldots, N$ **do**
    if z_t available **then**
      Sample $\widetilde{\theta}_t^i \sim p(\theta_t | \theta_{t-1}^i, \Sigma_\theta)$;
    else
      Set $\widetilde{\theta}_t^i = \theta_{t-1}^i$;
    end if
    Sample $\widetilde{x}_t^i \sim p(x_t \mid x_{t-1}^i; \widetilde{\theta}_t^i)$;
    Compute $\widetilde{w}_t^i = w_{t-1}^i p(z_t \mid \widetilde{x}_t^i; \widetilde{\theta}_t^i)$;
  end for
  Normalize $\widehat{w}_t^i = \widetilde{w}_t^i / \sum_j \widetilde{w}_t^j$, $i = 1, \ldots, N$;
  for $i = 1, \ldots, N$ **do** ▷ Resample
    Sample k from the set $\{1, \ldots, N\}$ with weights $\{\widehat{w}_t^j\}_{j=1}^N$;
    Set $x_t^i = \widetilde{x}_t^k$, $\theta_t^i = \widetilde{\theta}_t^k$, $w_t^i = 1/N$;
  end for
end for

3.3. Liu–West filter

A popular approach trying to avoid the artificial overdispersion of the parameter due to the stochastic evolution of the parameter has been proposed in (Liu & West, 2001), referred to here as "Liu–West filter". Their approach has been widely used due to two advantages: it is independent of the specific model and it is easy to implement. Examples of its use in prognostics application are e.g.(Hu, Baraldi, Di Maio, & Zio, 2015; Peng et al., 2018).

The Liu–West filter, similarly to the diffusive particle filter of Sec. 3.2, evolves the parameter in time with a stochastic process. Unlike it, the process is tuned such that the mean and covariance of the marginal parameter distribution stays invariant during the parameter update process. The overdispersion of the parameter estimation is therefore kept under control, avoiding the main drawback of the diffusive bootstrap filter.

At each time t, the parameter value for the i^{th} particle is sampled from the modified stochastic process

$$\theta_t^i \sim N(m_t^i, \Sigma_t),$$

with suitable values for m_t^i and Σ_t. To calculate these, the (weighted) mean m_t and covariance $\Sigma_{m,t}$ of the marginal distribution over all θ_{t-1}^i are determined. These are then used to get

$$m_t^i = a\theta_{t-1}^i + (1-a)m_t \tag{6}$$

and

$$\Sigma_t = (1-a^2)\Sigma_{m,t}, \tag{7}$$

In Eqs. (6) and (7), the scalar tuning parameter $a \in [0, 1]$ is used in both m_t^i and Σ_t such that the marginal distribution of the newly sampled $\{\theta_t^i\}_{i=1}^N$ have the same mean m_t and covariance Σ_t as before. In this way the overdispersion from the sampling of θ_t^i is kept at a minimum.

Two limiting cases can be seen: If the coefficient $a \to 1$ the parameter values of the particles do not move over a time step, the Liu–West filter approaches the static particle filter. Conversely if $a \to 0$ all particles have parameters drawn from a common normal distribution, independent of the individual parameter values of the particles at the previous time step. The parameter a is often chosen very close to $a \to 1$ (e.g. 0.995) or even adapted over time to cope with the improved knowledge of the parameter. The method was inspired by the analogy of the marginal parameter distribution in the diffusive update step with a kernel density estimation or a Gaussian mixture model centered around the m_t^i.

A possible implementation of the Liu–West filter is given in Algo. 4; the main difference is the modified calculation of the update of the parameter value of the particles. We give here only the simplest implementation, whereas in (Liu & West, 2001) some additional importance sampling steps are used in addition.

Algorithm 4 The Liu–West particle filter

Initialize $\{x_0^i, \theta_0^i, w_0^i = 1/N\}_{i=1}^N$;
for $t = 1, \ldots, T$ **do** ▷ Propagate
  for $i = 1, \ldots, N$ **do**
    Determine mean m_t and variance Σ_t of the marginal of the θ_{t-1}^i;
    Determine $m_t^i = a\theta_{t-1}^i + (1-a)m_t$;
    Sample $\widetilde{\theta}_t^i \sim \mathcal{N}(m_t^i, (1-a^2)\Sigma_t)$;
    Sample $\widetilde{x}_t^i \sim p(x_t | x_{t-1}^i, \widetilde{\theta}_t^i)$;
    Evaluate the corresponding weights $w_t^i \propto w_{t-1}^i p(z_t | x_t^i, \widetilde{\theta}_t^i)$;
  end for
  for $i = 1, \ldots, N$ **do** ▷ Resample
    Sample k from the set $\{1, \ldots, N\}$ with weights $\{\widehat{w}_t^j\}_{j=1}^N$;
    Set $x_t^i = \widetilde{x}_t^k$, $\theta_t^i = \widetilde{\theta}_t^k$, $w_t^i = 1/N$;
  end for
end for

The Liu–West filter often works in practice, but requires tuning of the hyperparameter a, which might be difficult to set to a reasonable value in a real application. A wrong value of a can lead, as before to an incorrect prediction.

Dealing with missing measurements can be done in two possible ways: The first one replaces $p(z_t \mid \cdots)$ by one and evolves the parameter θ_t^i in the same way as for time steps with measurements. The second one follows the proposal above and keeps the parameter fixed for that time step. This corresponds formally to choosing $a = 1$ for them.

4. STORVIK FILTER

In (Storvik, 2002; Johannes & Polson, 2006; Erol, Li, Ramsundar, & Russell, 2013) the authors propose a class of particle filter approaches that are exact with respect to the parameter distribution. Even though they slightly differ in their specific implementation, they are based on the same basic concept and we refer to them together as "Storvik filter".

The main problem with parameter estimation in SSMs, and therefore also in SMC, is the increasing number of measurement data and hidden states, which makes evolving the parameter distribution progressively harder over time. The Storvik filter assumes the existence of a finite-dimensional sufficient statistic $s(x, z)$ for the distribution of the parameter given the states $x_{0:t}$ and the measurements $z_{0:t}$. Denoting $s_t = s(x_{0:t}, z_{0:t})$, sufficiency means that

$$p(\theta \mid x_{0:t}, z_{0:t}) = p(\theta \mid s_t).$$

The value of s_t carries all the relevant information contained in the history of $x_{0:t}$ and $z_{0:t}$. In addition, the Storvik filter requires a recursive rule

$$s_t = S(s_{t-1}, x_t, z_t).$$

to update the sufficient statistic with each new state and measurement.

The existence of a sufficient statistic with a finite and fixed dimension independent of t is not guaranteed. The Fisher–Pitman–Koopman–Darmois theorem states that such a finite sufficient statistic s_t exists if and only if the distribution of θ belongs to the exponential family, see e.g. (Barankin & Maitra, 1963). This is often the case and thus guarantees a wide applicability of the Storvik filter. Especially many models use normal distributed process noise terms together with a linear dependency of the parameter, which can be addressed with this approach as discussed in (Erol et al., 2013). Extensions to distributions that are not members of the exponential family can be found in (Johannes & Polson, 2006), where the authors consider mixtures of exponential families for this case. Finally, in (Joyce & Marjoram, 2008) the authors discuss the determination of approximately sufficient statistics from data, if exact sufficient statistics are not available.

The Storvik filter is given in Algo. 5. It shares a number of similarities with the already discussed filters in that the parameter is evolved as well in each time step. The particles contain in addition to x_t^i and θ_t^i also s_t^i, which are used to sample new values of $\theta_t^i \sim p(\theta_t^i \mid s_t^i)$.

5. STORVIK FILTER WITH MISSING MEASUREMENTS

Incorporating missing measurements into the Storvik filter is not straightforward, as was in the other cases, due to the re-

Algorithm 5 The Storvik particle filter

1: Initialize $\{x_0^i\}_{i=1}^N$;
2: Compute $\{s_0^i = s(x_0^i, z_0)\}_{i=1}^N$;
3: **for** $t = 1, \ldots, T$ **do** ▷ Propagate
4: **for** $i = 1, \ldots, N$ **do**
5: Sample $\theta_t^i \sim p(\theta_t \mid s_{t-1}^i)$;
6: Sample $\widetilde{x}_t^i \sim p(x_t \mid x_{t-1}, \theta_t^i)$;
7: Evaluate $\widetilde{w}_t^i = w_{t-1}^i p(z_t \mid \widetilde{x}_t^i, \theta_t^i)$;
8: Compute $\widetilde{s}_t^i = S(s_{t-1}^i, \widetilde{x}_t^i, z_t)$;
9: **end for**
10: Normalize $\widehat{w}_t^i = \widetilde{w}_t^i / \sum_j \widetilde{w}_t^j$, $i = 1, \ldots, N$;
11: **for** $i = 1, \ldots, N$ **do** ▷ Resample
12: Sample k from the set $\{1, \ldots, N\}$ with weights $\{\widehat{w}_t^j\}_{j=1}^N$;
13: Set $x_t^i = \widetilde{x}_t^k$, $s_t^i = \widetilde{s}_t^k$, $w_t^i = 1/N$;
14: **end for**
15: **end for**

quired update of the sufficient statistic in each step. We have identified two possible approaches:

1. Resampling the parameter at each step, despite missing measurements. We denote this choice by "U" as in "Updating".
2. Freezing the parameter to the value at the last observed time. We denote this choice by "F" as in "Frozen".

The two approaches can be shown to result from splitting the joint posterior distribution for the evolution of the state from $t+1$ to $t+k$ without measurements into

$$p(x_{t+1:t+k}, \theta) = p(\theta | s_{t+k}) p(x_{t+1:t+k}),$$

showing that the sufficient statistic needs to be updated using all $x_{t+1:t+k}$ to get the correct distribution of θ at the end. The distribution $p(x_{t+1:t+k})$ on the other hand can be decomposed in two different ways, either as

$$p(x_{t+1:t+k}) = \int p(x_{t+k}|\theta) p(x_{t+k-1}|\theta) \ldots p(x_{t+1}|\theta)\, p(\theta | x_{1:t}, z_{1:t}) d\theta$$

which corresponds to sampling one θ at the last time step with a measurement and sampling all new values of x with it, which is the "F" approach. Alternatively one can write

$$p(x_{t+1:t+k}) = \int p(x_{t+k}|x_{t+k-1}, \theta) p(\theta | s_{t+k-1}) d\theta \times p(x_{t+1:t+k-1})$$

which corresponds to updating the sufficient statistic after each step and sampling a new θ from it. This is the "U" approach. This shows that both approaches are identical in principle, but could still be more or less efficient in applications.

A possible implementation of both approaches is given in Algos. 6 and 7 for the "U" and "F" approach replacing lines 3 to 9 in Algo. 5, respectively. Please note the similarity of the two approaches to the one previously discussed.

Algorithm 6 Storvik particle filter "U" for predictions

for $i = 1, \ldots, N$ **do**
Sample $\theta_t^i \sim p(\theta_t \mid s_{t-1}^i)$;
Sample $x_t^i \sim p(x_t \mid x_{t-1}, \theta_t^i)$
Set $w_t^i = w_{t-1}^i$
Compute $s_t^i = S(s_{t-1}^i, x_t^i)$
end for

Algorithm 7 Storvik particle filter "F" for predictions

Set $\theta_t^i = \theta_{t-1}^i$ for $i = 1, \ldots, N$
for $i = 1, \ldots, N$ **do**
Sample $x_t^i \sim p(x_t \mid x_{t-1}, \theta_t^i)$
Set $w_t^i = w_{t-1}^i$
Compute $s_t^i = S(s_{t-1}^i, x_t^i)$
end for

6. Simulation study

We compare the performances of the four algorithms (parameter-augmented bootstrap particle filter, diffusive bootstrap particle filter, Liu–West filter, and Storvik filter) when applied to two simple models mimicking typical degradation dynamics. In addition we use a MCMC implementation in order to get the exact posterior distribution for all cases, using the JAGS probabilistic programming language (Plummer, 2003).

6.1. The linear model

The simplest model is the one of a Brownian motion with drift for x_t

$$x_{t+1} \sim \mathcal{N}(x_t + \alpha, \sigma_x) \tag{8}$$

together with a normal distributed measurement error

$$z_t \sim \mathcal{N}(x_t, \sigma_z) \tag{9}$$

This model is also referred to in the literature as the Whitmore model (Whitmore, 1995). For this study, we assume that only α is unknown and that σ_x and σ_z are known, so that only α and x_t need to be determined. This model falls into the class of having normal distributed process noise and being linear in the parameter. Therefore the sufficient statistic is known to be the mean and standard deviation of the distribution of the parameter α.

The prior distributions are assumed to be given as

$$p(\alpha) = \mathcal{N}(\alpha_0, \sigma_{\alpha_0}) \tag{10}$$

and

$$p(x_0) = \mathcal{N}(z_0, \sigma_{x_0}) \tag{11}$$

centered around the first measurement z_0.

To simulate the data set we have used $x_0 = 1000$, $\alpha = -8.65$, $\sigma_x = 1$, and $\sigma_z = 5$. The prior distribution parameters are $\alpha_0 = \alpha = -8.65$, $\sigma_{\alpha_0} = \sqrt{5}$ and $\sigma_{x_0} = 10$.

6.2. The stretched exponential or Weibull model

For many devices, e.g. for batteries or capacitors, the degradation process accelerates over time. The Weibull function, also known as stretched exponential, is often employed to capture this. We use a model of the form

$$x_{t+1} \sim \mathcal{N}\left(x_t - 3\alpha\left(-\ln\left(\frac{x_t}{X}\right)\right)^{1-1/3} x_t, \sigma_x\right),$$

which is chosen to follow approximately a Weibull function with shape parameter 3, but in the form of a time-independent stochastic process. The measurement model reads as before

$$z_t \sim \mathcal{N}(x_t, \sigma_z),$$

We assume the same prior distribution for α and x_0 as in Eqs. (10) and (11).

As in the linear model, the process noise is modeled as normal distributed and the dependency on the parameter is linear, even if the evolution of x_t is not. This makes the sufficient statistic to be as before the mean and standard deviation of the distribution for α.

For the simulation we have used initial condition $x_0 = 995$, maximum value $X = 1000$, $\sigma_x = 1$, $\sigma_z = 5$, and $\alpha = 1/80$. For the prior we use $\alpha_0 = \alpha = 1/80$ and $\sigma_{\alpha_0} = 6 \cdot 10^{-3}$ and where $\sigma_{x_0} = 1$ to avoid sampling impossible values $x_0 \geq X$. Parameters were chosen, such that the two models are comparable in terms of the degradation path.

For all tests we simulate 100 time steps with the measurements thinned, such that only every 5$^{\text{th}}$ time step was recorded to verify how efficiently the methods can treat missing measurement. We also stopped the estimation phase at either time horizon $t_0 = 30$ or 75 and continued with the prediction part only. The data set used, as well as the mean degradation curve for the two models are shown in Fig. 1.

6.3. Results

Figure 2 demonstrates the issues that can affect the static and the diffusive particle filters and the Liu–West filter. Measurements are available until $t_0 = 75$, after which only predictions were done. Note that we selected hyperparameters in order to exaggerate the issues. A more fine-tuned approach would lead to better agreement with the MCMC results.

With respect to the MCMC based reference result, given in Figure 2(d) we observe:

- Figures 2(a) and (b) demonstrate the underestimation and overestimation of the parameter uncertainty when using the static parameter-augmented particle filter and the diffusive particle filter, respectively.
- Figure 2(c) demonstrates that also the Liu–West filter can yield overdispersed results, even if overall less severe than for the diffusive particle filter case.

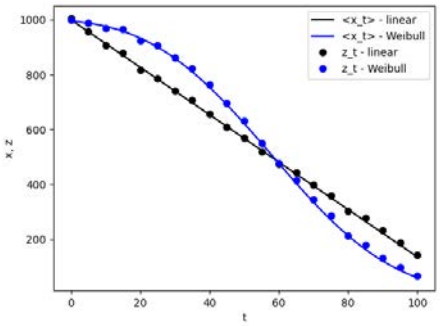
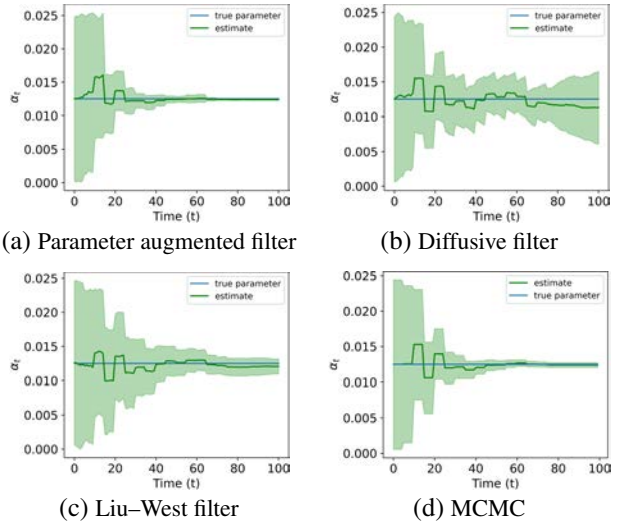


Figure 1. The mean value for x_t and the thinned measurement data z_t for the two models is shown.

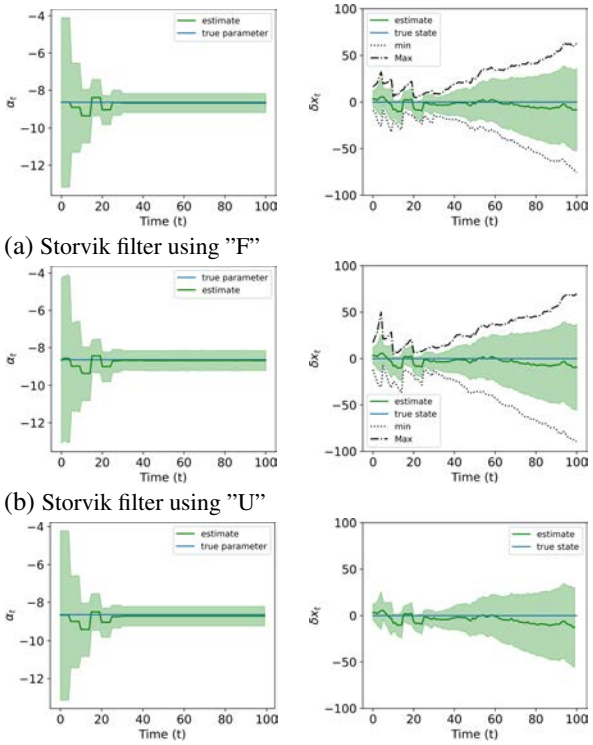


(a) Parameter augmented filter (b) Diffusive filter

(c) Liu–West filter (d) MCMC

Figure 2. Parameter estimation for the Weibull model for the three different particle filter: (a) static bootstrap, (b) diffusive, and (c) Liu–West filter. (d) gives the reference results using the MCMC approach. Mean values and uncertainty in terms of two standard deviations are given.

In Fig. 3 we give the results for the two implementations ("F" and "U") of the Storvik filter for missing measurements for the linear model, in Fig. 4 for the Weibull model. In order to focus on the performance of the two algorithm in the prediction phase, we set the time horizon to $t_0 = 30$. We observe that both implementations of the Storvik filter for missing measurements lead to results that are consistent with the MCMC result, despite (or due to) the absence of tuning hyperparameters. We also observe that the two implementations are practically indistinguishable, with the exception of a slightly more unstable behavior of the U implementation, visible in the initial time period of the Weibull model in

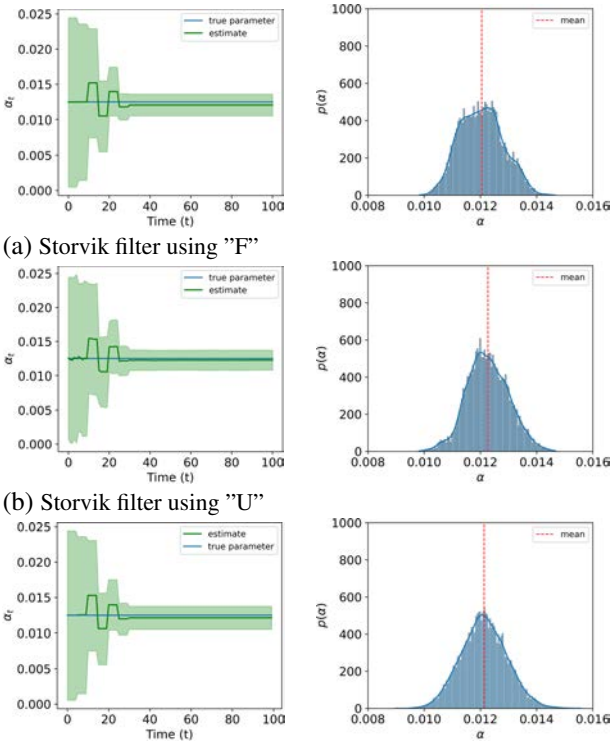


(a) Storvik filter using "F"

(b) Storvik filter using "U"

(c) MCMC

Figure 3. Parameter and state estimation of the linear model with the Storvik filter: Results using the two different treatments in case of missing data are compared using (a) the Frozen and (b) the Update approach. Results are compared with the reference result using the MCMC approach. On the left the evolution of the parameter estimation, on the right the difference between predicted and the true state is given. In both cases the mean and the uncertainty in terms of two standard deviation is shown.



(a) Storvik filter using "F"

(b) Storvik filter using "U"

(c) MCMC

Figure 4. Parameter estimation of the Weibull model with the Storvik filter: Results using the two different treatments in case of missing data are compared using (a) the Frozen and (b) the Update approach. Results are compared with the reference result using the MCMC approach. On the left the evolution of the parameter estimation is given showing the mean and the uncertainty in terms of two standard deviation, whereas on the right we give the full distribution for the final time step.

Fig. 4(b).

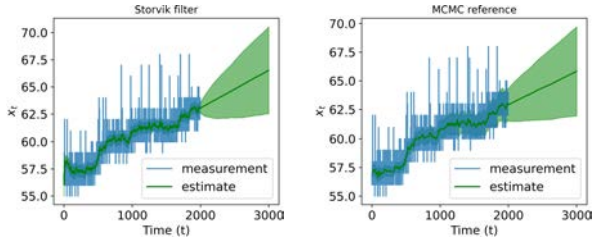
7. Application to real data with breaker opening times

In this section, we test the application of the Storvik filter against real data from an application with circuit breakers. In this case model misspecification is present and could undermine the applicability of the approach.

Circuit breakers are protection devices to interrupt short circuit currents occurring in an electric network. They are operated by mechanical mechanisms whose malfunction is one of dominant failure modes for them. The time required to open or close the contacts is the commonly monitored property. For instance, a reduction in a spring force or an increase in friction leads to an increase of this time. Hence, tracking it as a function of the number of operations enables to predict the end-of-life of these devices.

The evolution of the time x_t of the mechanical opening/closing operation is in general stochastic. Please note that t in this case typically refers to the number of operations performed instead of the time in operation. We describe it by the linear model as given in Eq. (8). The main issue with the data is that the measurement error is not following a normal distribution as assumed in Eq. (9). In fact, because of the signal processing performed during acquisition, the data is strongly quantized, as can be seen in Fig. 5(a). This was already discussed and analysed in (Hencken, 2021), which concluded that assuming normal distributed error gives reasonable results in a full analysis.

As circuit breaker often perform a larger number of operations, the use of a sequential approach will be an advantage in practice. We therefore explore here whether a particle based approach based on the Storvik filter is suitable. The model features three unknown parameters: the drift α and the two standard deviation for the process σ_x and the measurement σ_z, which leads to a more complex sufficient statistic. Following the usual normal-inverse-gamma model, the sufficient statistic consists of six variables, which are mean and standard deviation of the normal distribution of the drift α and



(a) Storvik filter (b) MCMC

Figure 5. Results of the application of the Storvik filter to the estimation of the degradation process of opening times x_t of a circuit breaker as a function of the number of oerations t. (a) shows the original data together with the state estimations and the prediction of the future evolution for the Storvik filter using the "U" method, (b) the same but for the MCMC reference approach.

the location and scale parameter of the inverse Gamma distribution for σ_x and similar for σ_z. For more details, we refer to (Storvik, 2002), where the sufficient statistic, as well as their update rules, are given.

As all measurements are available, we focus on a comparison of the basic Storvik filter without missing measurements with the exact result as given by the MCMC approach. The results of the state estimation and the expected future evolution of the two models are shown in Fig. 5. The Storvik filter is able to estimate the states quite similar to the ones found in the reference approach and in addition is able to capture the future evolution. Some slight deviations are visible, especially of the MCMC results showing a slightly larger uncertainty at the end of the measurements. But this demonstrates in a first step the possible application of the Storvik filter in real applications.

8. Conclusions and Outlook

Model-based prognostics requires joint state and parameter estimation. A sequential approach is most suitable to avoid increase in computational complexity over time. Several approaches involving particle filters and their potential issues have been discussed. We have focused also on the need of a robust treatment of time steps with missing measurements either due to irregular data acquisition or for the predictions needed for the RUL calculation. We have explored the use of the Storvik filter for prognostics application as an exact parameter estimation approach. We have shown that it can be naturally extended to incorporate missing measurements in two ways, which are similar to the ones discussed for the other particle filter approaches. Its main limitation is that it is restricted to problems allowing for the existence of a sufficient statistic. Simulations using two simple models showed the robustness and reliability of the Storvik filter, whereas we demonstrated as well, that other approaches can lead to erroneous results. We have also applied it to one real world examples, in order to test its applicability in a case, where the assumed model is not valid.

Prognostics using particle filters is an active area of research and development of real applications. The promising results with the Storvik filter should be further explored and its applicability to more complex problems, including higher dimensional state space and parameter vectors, but also to models beyond the restricted class studied here, should be explored. Finding suitable sufficient statistics in these more general models, even outside the exponential family, is another line of research to be undertaken.

References

An, D., Choi, J.-H., & Kim, N. H. (2013, July). Prognostics 101: A tutorial for particle filter-based prognostics algorithm using Matlab. *Reliability Engineering & System Safety*, *115*, 161–169. doi: 10.1016/j.ress.2013.02.019

Barankin, E. W., & Maitra, A. P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, 217–244.

Chopin, N., & Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-47845-2

Doucet, A., de Freitas, N., & Gordon, N. (2001). An introduction to Sequential Monte Carlo methods. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential monte carlo methods in practice* (pp. 3–14). New York, NY: Springer New York.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On Sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*, 197 – 208.

Erol, Y., Li, L., Ramsundar, B., & Russell, S. J. (2013). *The extended parameter filter.*

Galar, D., Goebel, K., Sandborn, P., & Kumar, U. (2021). *Prognostics and remaining useful life (rul) estimation: Predicting with confidence*. CRC Press.

Goebel, K., Daigle, M., Saxena, A., Sankararaman, S., Roychoudhury, I., & Celaya, J. (2017). *Prognostics: The science of making predictions*. Createspace Independent Pub.

Hencken, K. (2021). Parameter estimation of a Wiener process of mechanical degradation through censored measurement of timings. *Proceedings of the 31st European Safety and Reliability Conference*, 317-324.

Hu, Y., Baraldi, P., Di Maio, F., & Zio, E. (2015). A particle filtering and kernel smoothing-based approach for new design component prognostics. *Reliability Engineering & System Safety*, *134*, 19-31. doi: https://doi.org/10.1016/j.ress.2014.10.003

Johannes, M. S., & Polson, N. G. (2006). Exact particle filtering and parameter learning..

Jouin, M., Gouriveau, R., Hissel, D., Péra, M.-C., & Zerhouni, N. (2016). Particle filter-based prognostics: Review, discussion and perspectives. *Mechanical Systems and Signal Processing*, *72*, 2–31.

Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, *7*(1).

Liu, J., & West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential monte carlo methods in practice* (pp. 197–223). New York, NY: Springer New York.

Peng, W., Ye, Z.-S., & Chen, N. (2018). Joint online rul prediction for multivariate deteriorating systems. *IEEE Transactions on Industrial Informatics*, *15*(5), 2870–2878.

Plummer, M. (2003, 04). JAGS: a program for analysis of bayesian graphical models using gibbs sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, *124*.

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011, August). Remaining useful life estimation – A review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14. doi: 10.1016/j.ejor.2010.11.018

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, *50*(2), 281-289. doi: 10.1109/78.978383

Whitmore, G. A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error. *Lifetime data analysis*, *1*(3), 307–319.

BIOGRAPHIES

Kai Hencken is corporate research fellow at ABB corporate research. He obtained a PhD in Theoretical Physics from the University of Basel in 1994. He was a post-doc at the University of Washington from 1995 to 1997 and at the University of Basel from 1997 to 2005, where he received his habilitation in 2000 and is a lecturer since. In 2005 he joined the theoretical Physics group at ABB corporate research. His research interests include the combination of physical modeling with statistical methods to solve problems related to industrial devices, as well as developing diagnostics and prognostics approaches.

Arthur Serres studied mathematics in EPFL, Imperial College London, and ETHZ. He is currently enrolled in the final year of a Master program at ETH, where he focuses on graph neural networks for transaction monitoring. He worked as a research intern in ABB corporate research and he is now a quantitative researcher. His interests include theoretical machine learning and probability theory.

Giacomo Garegnani obtained a PhD in Mathematics from EPFL in 2021, with a thesis on inverse problems involving partial and stochastic differential equations, and on probabilistic numerical methods. He is now a scientist at ABB corporate research. His research interests include uncertainty quantification of numerical solvers, identifiability of differential models, and methods of statistical inference for condition monitoring.

PHM for Spacecraft Propulsion Systems: Developing Resilient Models for Real-World Challenges

Takanobu Minami[1], Dai-Yan Ji[2] and Jay Lee[3]

[1,2,3]*Center for Industrial Artificial Intelligence, Department of Mechanical Engineering, University of Maryland, MD, USA*
minamitu@umd.edu, jidn@umd.edu, leejay@umd.edu

[1]*Komatsu Ltd., Tokyo, Japan*

ABSTRACT

This paper extends the research presented at the Prognostics and Health Management (PHM) Asia-Pacific 2023 Conference Data Challenge, focusing on a more pragmatic approach to spacecraft propulsion system health assessment. While the previous competition saw a variety of solutions, they predominantly relied on the assumption of highly stable initial hydraulic conditions – an idealization seldom met in real-world scenarios. In practical settings, factors such as operational noise, recent operational states, and ambient environmental conditions significantly disrupt this stability, rendering such solutions less feasible. Addressing this gap, our current study introduces a novel diagnostic model capable of valve faults without depending on the initial stable state of hydraulics. This approach marks a significant shift from our previous methodology, which primarily utilized similarity measures and physics-inspired features to classify health states and identify solenoid valve faults in spacecraft propulsion systems. The proposed model in this paper is validated against a diverse set of conditions, emphasizing its robustness and applicability in fluctuating real-world scenarios. Our findings demonstrate that the new model not only effectively diagnoses system health under varied and less controlled conditions but also enhances the practicality of spacecraft health management, offering a more adaptable solution in the face of operational uncertainties.

1. INTRODUCTION

Propulsion systems in spacecraft are essential for navigating through the cosmos, and their dependable and effective operation is critical. Therefore, the health management of these systems is of utmost significance. The role of Prognostics and Health Management (PHM) is central in ensuring this dependability, as it allows for the early identification and assessment of potential problems or irregularities within the propulsion mechanisms.

To promote Spacecraft PHM, the Japan Aerospace Exploration Agency (JAXA) created and released a dataset to the public (Tominaga et al., 2023), and at the same time, a data challenge was held at the PHM Asia Pacific 24 conference to facilitate the use of this data (PHMAP 2023 Secretariat, 2023). The Data Challenge required complex diagnostics such as analytical detection, classification, and regression, and many institutions took on the challenge. Despite the complexity of the problem, the top three teams of the data challenge ultimately succeeded in creating highly accurate models, and these results have been compiled and published in papers (Kato, et al., 2023) (Lee et al., 2023) (Minami & Lee, 2023). This effort was an important step in the promotion of spacecraft PHM. However, there are two major problems in adapting these models to the real world.

The first problem is the presence of non-noise regions that are unique to this data set. All of the top three teams found and used a time region in the given pressure sensor data that is completely free of noise. In this time region, all data sets with identical health conditions have the same pressure values, and the differences among Spacecraft individuals and data are zero. Specifically, the given pressure time series data is completely free of noise/variation in the initial 0.1 seconds (0 to 0.1 sec) of the 1.2 seconds. This is evidenced by the results of the data analysis (Kato, et al., 2023). This is presumably because this data set was generated by simulation. Since this specificity is considered to be different from the behavior of pressure in the real world, there is a concern that even if a high-performance model is created using only the completely noise-free portion of this data set, it will be completely useless in the real world if any noise is added, or if there is any variation in the data. To dispel this concern, it is necessary to evaluate the model using data with noise/variance.

The second problem is the use of valve open-state data. The data given are data from three iterations of valve opening and

closing, and the three proposed models use only the data from that first valve open state. However, in the valve open state, the propulsion system is not closed as a system and its pressure behavior is subject to external influences. Therefore, the pressure in the valve open state is a very complex and unpredictable behavior. Therefore, in pursuit of a robust model, state estimation is required using the pressure in the valve closed state, i.e., when the system is closed.

Because of these two major problems, the models proposed in the data challenge may not stand up to use in the real world.

To summarize the above points:

Firstly, regarding the variance-free region of the dataset:

- The spacecraft valve dataset contains an unrealistic time region that is free of variance.
- The models proposed in the data challenge use this variance-free time region, which may result in poor performance when applied to real-world data.
- To ensure the proposed models perform well in the real world, it is necessary to validate them using time regions with variance.

Next, regarding the data during valve opening and closing:

- When the valve is open, the system is open, making the sensor data complex and unpredictable. This cannot be verified until tested with real-world data.
- For building a robust model, it is preferable to use data from the closed system when the valve is closed.
- All models proposed in the data challenge are designed and trained using data from the valve-open state.
- To construct a robust model, it is necessary to design new models based on data from the valve-closed state.

To address this issue, this paper examines and evaluates the models for the PHM of spacecraft valve under the restriction that data from the variance-free portion is not used and assumes following two cases: Case 1 uses data from the valve open state, while Case 2 uses data from the valve closed state.

Model construction was examined under these scenarios to promote the construction of a more robust PHM model.

2. PROBLEM STATEMENT

The PHM Asia-Pacific 2023 Conference Data Challenge focused on Prognostics and Health Management for spacecraft propulsion systems, with the system's schematic illustrated in Figure 1. The training dataset provides 177 sets of synthetic data produced by simulations. Each set includes measurements from seven pressure sensors labeled P1 to P7, as depicted in Figure 1. These measurements were taken at a sampling rate of 1 kHz, throughout 1200 ms, and encompass three cycles of valve open-close operations, as shown in Figure 2.

The training dataset covers three distinct spacecraft, labeled #1 through #3, and it encompasses three different health conditions: normal operation, bubble anomalies, and solenoid valve faults. Solenoid valve faults could potentially occur in one of the four valves labeled SV1 through SV4, as shown in Figure 1. In the event of a fault, the solenoid valves may open anywhere from 0% to 100% of their full range. Under normal conditions, they open 100%. Note that the training data only include cases in which the valve open ratios are 0%, 25%, 50%, 75%, and 100%. The competition aims to utilize the 177 training data points to evaluate the health conditions of the 46 test data points. Half of the test data originates from spacecraft #4, which is not represented in the training set.

In this study, we focus only on the most complex task of estimating valve apertures. Two problem settings, Case 1 and Case 2, are used to validate the model for the two major problems described in the Introduction. Each is described in detail in the following sections.

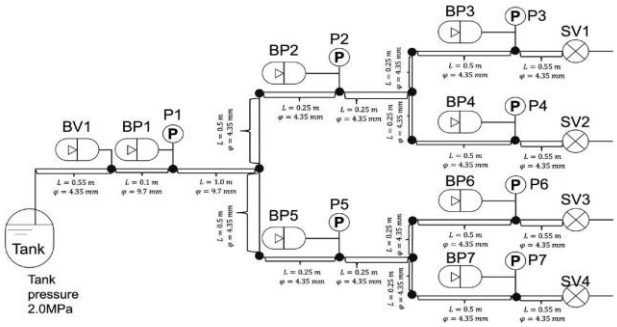


Figure 1. Schematic of experimental propulsion system

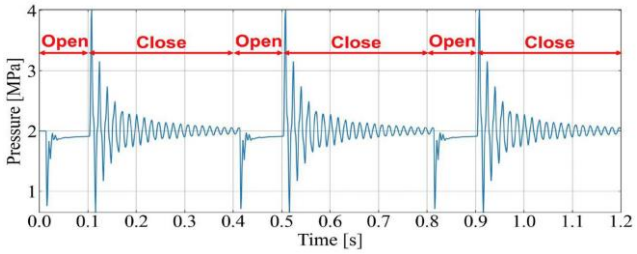


Figure 2. Typical pressure profile

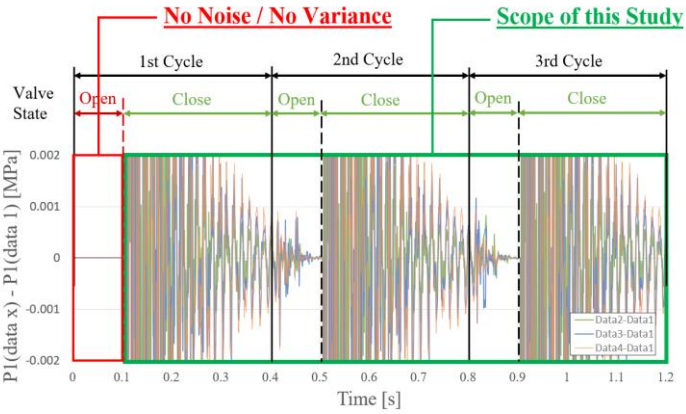


Figure 3. Pressure differences among normal data

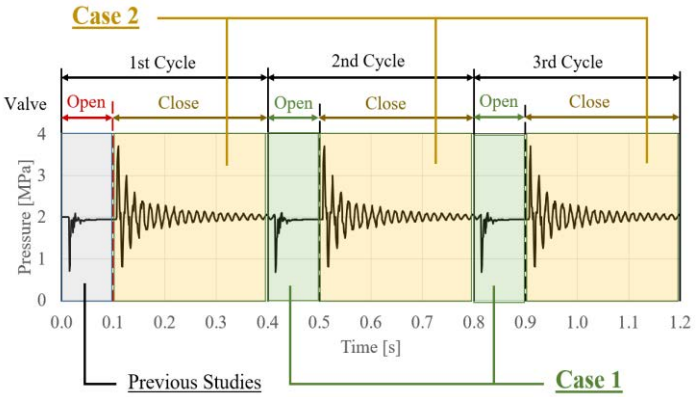


Figure 4. Data Regime

2.1. Case 1: Valve opening ratio prediction using data at valve opening with noise/variance

In Case 1, only the data of the valve open state of the second and third cycle of the valve open/close cycles is used. Specifically, as shown in the green area of Figure 4, out of the total 1.2 seconds of pressure data, only 0.4 to 0.5 seconds and 0.8 to 0.9 seconds, for a total of 0.2 seconds of data are used.

Case 1 evaluates model performance with the following metrics as well as data challenge

The evaluation metric is as follows:

$$Total\ Score = \frac{\sum_i^{Ntest} Score_i}{\sum_i^{Ntest} Score(max)_i} \quad (1)$$

Here, N_{test} is the number of test data. $Score_i$ is as follow:

$Score_i$: For the solenoid valve correctly identified as fault, prediction of the opening ratio: max (20-|truth – prediction|, 0)

For spacecraft #4, $Score_i$ are doubled, considering the difficulty. $Score(max)$ is the score if there were no prediction errors. Therefore, the total score can range from 0% to 100%.

2.2. Case 2: Valve opening ratio prediction using data at valve closed

In Case 2, only the data of the valve closed state for the 1st, 2nd, and 3rd cycles of the whole sensor data are used. Specifically, as shown in the orange area of Figure 4, out of a total of 1.2 seconds of pressure data, 0.1 to 0.4, 0.5 to 0.8, and 0.9 to 1.2 seconds of pressure data are used.

Since Case 2 is more difficult than Case 1 and it is difficult to estimate the valve opening ratio with continuous values, set the classes according to the valve opening ratio as shown in Figure 5, and set the problem as a classification problem to predict the valve opening ratio class instead of a regression problem to predict the numerical value of the valve opening ratio.

The classification models are evaluated using the following metric where TP is the total number of test data that are correctly classified.

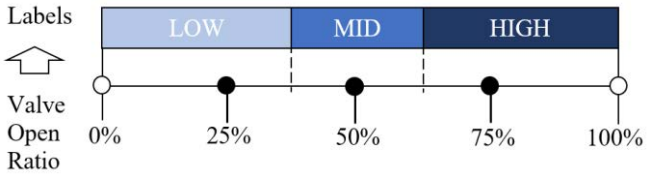


Figure 5. Labeling of valve opening ratio

$$Classification\ Accuracy = \frac{TP}{N_{test}} \quad (2)$$

3. Background: Model Selection

PHM often face the challenge of dealing with noise and variability in data, which can obscure the fundamental patterns necessary for accurate diagnosis and prediction. A common approach to address this issue is the use of filtering techniques, such as moving averages and other signal processing methods.

Simple filtering techniques, such as moving averages and other basic smoothing methods, are widely used in PHM to reduce noise and improve signal quality. For instance, Mubarak et al. (2023) demonstrated that applying a moving average filter to time series signals outperformed traditional condition monitoring methods in tasks such as Rolling Element Bearing Fault Diagnosis and Hydraulic Accumulator State Classification. Similarly, Boškoski and Urevc (2011) showed that passing vibration signals through a band-pass filter effectively removed noise, enhancing the accuracy of bearing fault detection by analyzing the envelope spectrum of the filtered signals.

However, these simple methods have significant limitations. The primary concern is their inability to distinguish between noise and useful information. As a result, essential diagnostic information may be inadvertently removed along with the noise. This is particularly problematic in scenarios like predicting valve opening degrees, where minute pressure fluctuations carry significant diagnostic value. Standard noise removal techniques are likely inappropriate here, as they can degrade model performance by losing critical diagnostic information.

To address the limitations of simple filtering, advanced techniques such as deep learning are utilized (Najafabadi et al., 2015). Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), excel in distinguishing noise from useful signals. For example, Baptista and Henriques (2022) used a one-

dimensional denoising GAN (1D-DGAN) to filter noise from turbofan engine operational data, significantly improving fault detection accuracy. Liu et al. (2018) proposed a novel bearing fault diagnosis method using an autoencoder in the form of an RNN. This method employed a Gated Recurrent Unit (GRU)-based denoising autoencoder to predict multiple vibration values of rolling bearings for the next period from the previous period. The proposed method demonstrated satisfactory performance with high robustness and classification accuracy.

While these advanced methods are effective in removing noise without losing useful information, they typically require large amounts of training data. Such extensive datasets are essential for the model to learn complex patterns and distinguish subtle differences between noise and useful signals.

In contrast, the dataset used in this study is very limited. This limitation makes the application of deep learning approaches impractical, as the model is likely to overfit the small dataset and fail to generalize to unseen data.

Given these constraints, rule-based models or simple models with fewer parameters are more suitable for this study. Therefore, this study focuses on designing and validating methods for the two cases set in the previous section, based on models proposed in the data challenge that meet these criteria. Specifically, we use the polynomial regression model based on pressure drop proposed by Minami and Lee (2023) and the similarity-based regression model proposed by Kato et al. (2023).

In Case 1, we directly utilize the existing models proposed in the data challenge to evaluate their performance in the presence of noise and variability. The primary focus is to assess whether and to what extent the performance of the previously proposed models degrades with increased variability.

In Case 2, since the models proposed in the data challenge are based on the assumption of valve open states, they cannot be used directly. This study examines the adaptation of these models' features to valve closed states. By doing so, it becomes possible to leverage the existing model structures while adapting them to new conditions.

4. METHODOLOGY AND RESULTS

In this section, the design and validation of models for two distinct cases are conducted. For both cases, a linear regression model is adopted as the benchmark method. This benchmarking methodology involves extracting nine types of basic statistics (Mean, Standard Deviation, Minimum, 25th Percentile, Median, 75th Percentile, Maximum, Skewness, Kurtosis) from each of the seven sensors. After extraction, dimensionality reduction is performed using PCA.

4.1. Case 1

In Case 1, the green area in Figure 4. Here, we examine how well the solution proposed in the data challenge maintains performance in a noisy and varied environment.

4.1.1. Methodology

As shown in Figure 1, there were two main valve opening prediction models implemented in the data challenge: one is the method that estimates the valve opening ratio by performing a polynomial fit based on the pressure drop/slope immediately after valve opening (Lee et al., 2023) (Minami & Lee, 2023). The other is the method that uses the similarity of the overall pressure during the first 0.1 seconds after the valve opens to estimate the pressure. (Kato, et al., 2023).

To adapt these proposed methods for Case 1, here, the predicted valve open ratio is calculated for each of the predictions for the 2nd cycle data (0.4 to 0.5 sec) and the 3rd cycle data (0.8 to 0.9 sec), and take the average of these is the final predicted value

4.1.2. Results

The prediction results from each model are shown in Table 1, and the calculation results of the estimation accuracy are shown in Figure 6.

Polynomial Fit's model is still able to maintain a high accuracy rate of 96%, albeit with lower accuracy, relative to previous results in the noiseless region. This suggests that the pressure drop is an important indicator that is not easily affected by noise. On the other hand, the model using Similarity shows a significant drop in accuracy, from 89% to 48%. This indicates that Similarity is susceptible to noise and has poor generalization performance when the valve is open. These results indicate that the Polynomial Fit method, which focuses on the initial pressure drop, is effective in estimating the valve opening ratio, even with noise and variation, as long as data on the valve opening state is available. On the other hand, since the data is a simulation and the number of N is small, it is necessary to verify the validity of this finding by measuring data in a setting closer to reality.

Table 1. Models and predicted valve opening ratio

Spacecraft	Case	Valve	Open Ratio [%]						
			Ground Truth	Estimation					
				Benchmark		①Polynomial		②Similarity	
				1st Cycle	**2nd, 3rd Cycle**	1st Cycle	**2nd, 3rd Cycle**	1st Cycle	**2nd, 3rd Cycle**
1	179	2	22	25	**40**	22	**22**	24	**24**
1	181	4	76	87	**72**	76	**71**	77	**79**
1	188	1	5	0	**0**	5	**4**	20	**19**
1	190	3	46	25	**42**	46	**46**	46	**42**
1	199	1	98	95	**100**	98	**99**	97	**96**
4	202	3	44	28	**41**	44	**44**	44	**42**
4	205	2	94	94	**90**	94	**95**	95	**37**
4	211	1	95	93	**95**	95	**95**	93	**62**
4	212	2	70	78	**82**	70	**71**	67	**16**
4	214	4	24	28	**20**	24	**25**	25	**21**

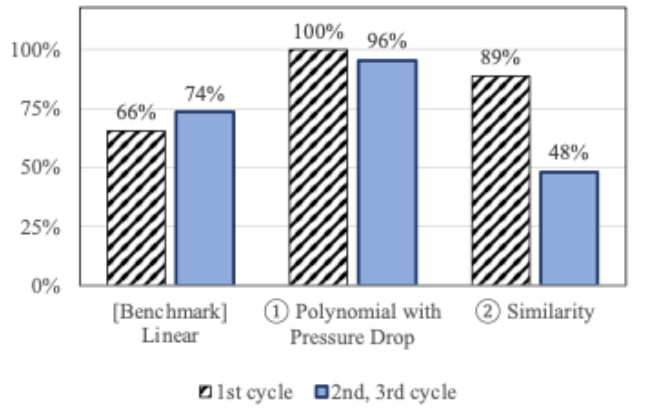


Figure 6. Regression accuracy

4.2. Case 2

In Case 2, only the data in the orange region of Figure 4 is used as a more practical but more difficult setting compared to Case 1.

4.2.1. Methodology

Unlike Case 1, the solution proposed in the data challenge cannot be used, thus a new model must be devised. Theoretically, the difference in pressure behavior with valve opening is determined only by the pressure state immediately before closing the valve, and it all returns to a constant pressure with time after closing. In other words, the difference in valve opening ratio has the greatest effect on the pressure immediately after the valve is closed, and as time passes, the difference in valve opening ratio has less effect on the pressure difference. Therefore, we devised the following two models that focus on the pressure behavior immediately after the valve is closed.

4.2.2. Method 1: Valve closing pressure surge

The first proposed model focuses on the pressure increase immediately after valve closing, similar to the focus on pressure drop in Case 1. As an example, shown in Figure 7, the pressure rise after valve closing is divided by the valve opening %, which may be used to classify the pressure rise. The label of the training data with the closest pressure based on the pressure after the specified time after the valve is closed is estimated as the label of the test data. Three models are created based on the pressure at 106 ms, 107 ms, and 108 ms after the valve was closed.

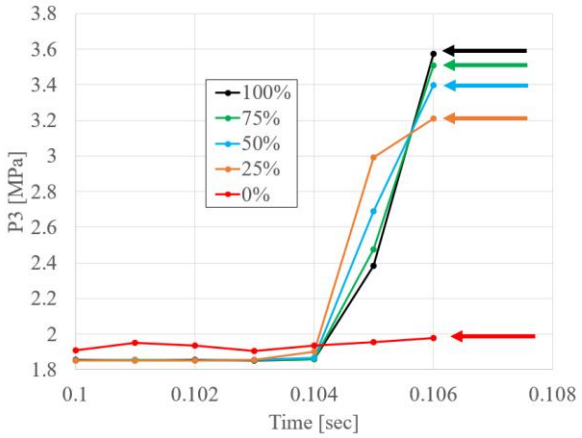


Figure 7. Example of pressure surge after valve closed (SV1)

$$Predicted\ Label = Label\left(\underset{i}{\operatorname{argmin}} \|P(train_i) - P(test)\|\right) \quad (3)$$

where training is the with training data, P is the pressure at the valve fault location, Label is the label of the training data, and Predicted Label is the label of the test data.

4.2.3. Method 2: Similarity

In this proposed model, as shown in Figure 8, the Euclidean distance is measured as the similarity of waveforms during a certain number of seconds after the start of valve closing operation, and the training data label with the highest similarity is used as the prediction label. Three models were created based on waveforms of different lengths (100 ms, 10 ms, and 5 ms).

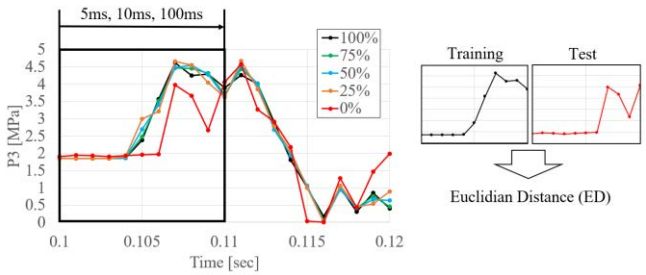


Figure 8. Similarity measurement process

$$Predicted\ Label = Label\ (\underset{i}{\text{argmin}}(\ ED(train_i, test))) \quad (4)$$

$ED(train_i, test)$ calculates the Euclidian distance of i-th training data and test data as a similarity.

4.3. Results

Table 2, Figure 9, and Figure 10 show the classification results and the results of valve opening estimation for the two models.

Table 2. Classification Results

Spacecraft	Case	Valve	Open Ratio [%]							
			Ground Truth	Estimation						
				Bench mark	①Polynomial			②Similarity		
					106 ms	107 ms	108 ms	100 ms	10 ms	5 ms
1	179	2	Low	Mid	Low	Low	Low	Low	Low	Low
1	181	4	High	High	Mid	High	High	High	High	High
1	188	1	Low	Low	Low	Low	Mid	Low	Low	Low
1	190	3	Mid	Mid	Mid	High	Mid	Mid	Mid	Mid
1	199	1	High	High	High	Mid	High	High	High	High
4	202	3	Mid	Mid	Mid	Mid	Low	High	Mid	Low
4	205	2	High	High	Low	High	Low	High	High	High
4	211	1	High	High	High	Low	Low	High	High	High
4	212	2	High	High	Low	High	Low	High	High	High
4	214	4	Low	Low	Low	Low	Low	High	Low	Low
No. of Correct Answer				9	7	7	5	8	10	9

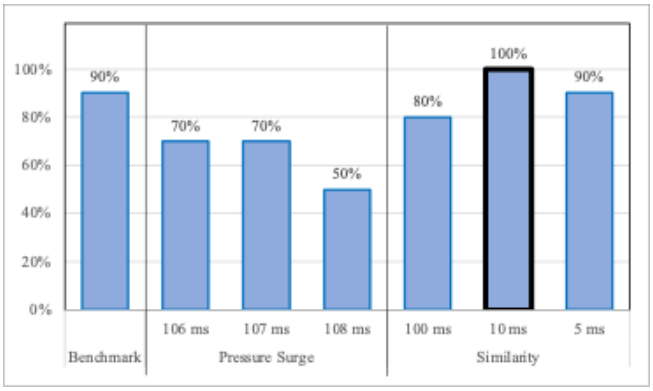


Figure 9. Classification Accuracy

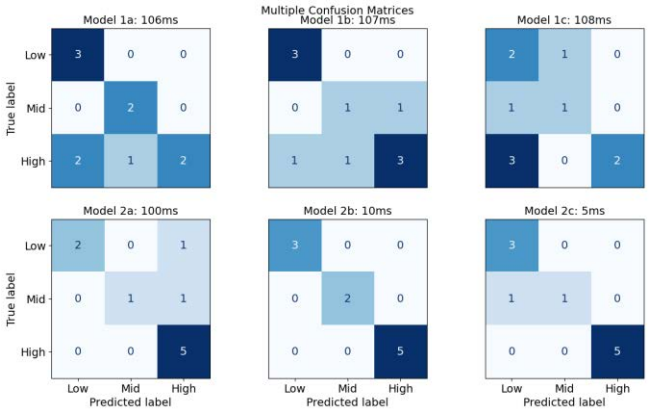


Figure 10. Confusion Matrix

The results show that Method 2 has better overall accuracy than Method 1 and the benchmarking method. The results of Method 1 show that in Case 1, high accuracy can be obtained only with pressure drop, while in Case 2, accuracy is not as good as it was, only with pressure rise. Possible reasons for this include variations in the timing of valve switching and the fact that the pressure rise is more complex than the pressure drop because it is caused by pressure propagation throughout the system.

Among the Method 2, the best accuracy was found when 10 ms waveforms were used. This suggests that there may be information useful for valve opening prediction in a specific interval. Although it is difficult to conduct a detailed analysis here due to the small amount of data, if more data were available, it would be possible to conduct an EDA and analyze the useful data areas.

From the above analysis, it is found that it is possible to use similarity to classify valve opening ratio classes and estimate intervals using only data for the closed valve state.

5. CONCLUSIONS

To construct a practical and robust spacecraft PHM model, we built and validated a valve opening prediction model with the constraint of eliminating noise/variation-free regions from the data set.

In Case 1, we verified the capability of the model proposed in the data challenge based on the valve opening data. The results showed that the regression model focusing on pressure drop had a regression accuracy of 96% even in the presence of noise and variability. On the other hand, the model using similarity was found to be only 48% accurate. This shows that the pressure drop model can produce robust results even with noise.

In Case 2, the model is built using only data from a closed system and closed valves. The model focusing on the pressure increase achieved only 70% accuracy in classification, while the model focusing on similarity achieved 100% accuracy. Further development of the model is needed to realize point estimation by regression rather than interval estimation of valve opening ratio by classification.

6. FURTHER RESEARCH

In this study, considering that system behavior generally becomes unstable when the system is open, a method that does not use the valve-open data from the dataset was proposed in Case 2. However, since the extent of instability depends on the application and the usage environment, it is necessary to collect data through experiments and verify the validity in future work.

REFERENCES

Baptista, M. L., & Henriques, E. M. (2022). 1D-DGAN-PHM: A 1-D denoising GAN for Prognostics and Health Management with an application to turbofan. Applied Soft Computing, 131, 109785.

Boškoski, P., & Urevc, A. (2011). Bearing fault detection with application to PHM Data Challenge. International Journal of Prognostics and Health Management Volume 2 (color), 32.

Kato, Y., Kato, T., & Tanaka, T. (2023, September). Anomaly detection in spacecraft propulsion system using time series classification based on k-nn. In PHM Society Asia-Pacific Conference (Vol. 4, No. 1).

Lee, S. K., Lee, J., Lee, S., Kim, B., Kim, Y. C., Lee, J., & Youn, B. D. (2023, September). Hybrid approach of xgboost and rule-based model for fault detection and severity estimation in spacecraft propulsion system. In PHM Society Asia-Pacific Conference (Vol. 4, No. 1).

Liu, H., Zhou, J., Zheng, Y., Jiang, W., & Zhang, Y. (2018). Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. ISA transactions, 77, 167-178.

Minami, T., & Lee, J. (2023, September). Phm for spacecraft propulsion systems: Similarity-based model and physics-inspired features. In PHM Society Asia-Pacific Conference (Vol. 4, No. 1).

Mubarak, A., Asmelash, M., Azhari, A., Haggos, F. Y., & Mulubrhan, F. (2023). Machine health management system using moving average feature with bidirectional long-short term memory. Journal of Computing and Information Science in Engineering, 23(3), 031002.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of big data, 2, 1-21.

PHMAP 2023 Secretariat. PHM Asia Pacific 2023 Conference Data Challenge. (2023, Aug 4). https://phmap.jp/program-data/

Tominaga, K., Daimon, Y., Toyama, M., Adachi, K., Tsutsumi, S., Omata, N., & Nagata, T. (2023, September). Dataset generation based on 1D-CAE modeling for fault diagnostics in a spacecraft propulsion system. In PHM Society Asia-Pacific Conference (Vol. 4, No. 1).

Takanobu Minami received his B.S. and M.S. degrees in mechanical engineering from Kyoto University in 2008 and in 2011, respectively. Currently, he is pursuing his Ph.D. degree in mechanical engineering with the University of Maryland, College Park, MD, USA, and is employed as an engineer in Komatsu Ltd. His research interests include machine learning, deep learning, prognostics and health management, and industrial AI.

Dai-Yan Ji received his B.S. degree in Electronic Engineering and M.S. degree in Communications Engineering from Feng Chia University, Taiwan, in 2009 and 2012, respectively. He is currently pursuing a Ph.D. degree in Mechanical Engineering with the University of Maryland, College Park, MD, USA. He was a Research Associate at the National Central University, Taiwan, in 2017. His research interests include machine learning, prognostics, and health management.

Jay Lee is Clark Distinguished Professor and Director of the Industrial AI Center in the Mechanical Engineering Dept. of the Univ. of Maryland College Park. His research is focused on intelligent analytics of complex systems including highly-connected industrial systems including energy, manufacturing, healthcare/medical, etc. He has been working with medical school in Traumatic Brain Injury (TBI) using multi-dimension data for predictive assessment of patient in ICU with funding from NIH and NSF. Previously, he served as an Ohio Eminent Scholar, L.W. Scott Alter Chair and Univ. Distinguished Professor at Univ. of Cincinnati. He was Founding Director of National Science Foundation (NSF) Industry/University Cooperative Research Center (I/UCRC) on Intelligent Maintenance Systems during 2001-2019. IMS Center pioneered industrial Ai-augmented prognostics technologies for highly-connected industrial systems and has developed research memberships with over 100 global company since 2000 and was selected as the most economically impactful I/UCRC in the NSF Economic Impact Study Report in 2012. He is also the Founding Director of Industrial AI Center. He was on leave from UC to serve as Vice Chairman and Board Member for Foxconn Technology Group (ranked 26th in Global Fortune 500) during 2019-2021 to lead the development of Foxconn Wisconsin Science Park (~$1B investment) in Mt. Pleasant, WI. In addition, he advised Foxconn business units to successfully receive five WEF Lighthouse Factory Awards since 2019. He is a member of Global Future Council on Advanced Manufacturing and Production of the World Economics Council (WEF), a member of Board of Governors of the Manufacturing Executive Leadership Council of National Association of Manufacturers (NAM), Board of Trustees of MTConnect, as well as a senior advisor to McKinsey. Previously, he served as Director for Product Development and Manufacturing at United Technologies Research Center (now Raytheon Technologies Research Center) as well as Program Director for a number of programs at NSF. He was selected as 30 Visionaries in Smart Manufacturing in by SME in Jan. 2016 and 20 most influential professors in Smart Manufacturing in June 2020, SME Eli Whitney Productivity Award and SME/NAMRC S.M. Wu Research Implementation Award in 2022

Probabilistic Uncertainty-Aware Decision Fusion of Neural Network for Bearing Fault Diagnosis

Atabak Mostafavi[1,2], Mohammad Siami[3], Andreas Friedmann[1], Tomasz Barszcz[4], Radoslaw Zimroz[5]

[1] *Fraunhofer Institute for Structural Durability and System Reliability LBF, Darmstadt, Hessen, 64289, Germany*
atabak.mostafavi@lbf.fraunhofer.de
andreas.friedmann@lbf.fraunhofer.de

[2]*Technische Universität Darmstadt, Department of Mechanical Engineering, Darmstadt, Hessen, 64287, Germany*
atabak.mostafavi @stud.tu-darmstadt.de

[3]*AMC Vibro Sp. z o.o., Pilotow 2e, 31-462, Kraków, Poland*
msiami@amcvibro.com

[4]*Faculty of Mechanical Engineering and Robotics, AGH University, Al. Mickiewicza 30, 30-059, Kraków, Poland*
tbarszcz@agh.edu.pl

[5]*Faculty of Geoengineering, Mining and Geology, Wroclaw University of Science and Technology, Na Grobli 15, 50-421, Wroclaw, Poland*
radoslaw.zimroz@pwr.edu.pl

ABSTRACT

Reliability is a central aspect of machine learning applications, especially in fault diagnosis systems, where only an accurate and reliable diagnosis system is economically justifiable, considering that any false diagnosis would lead to an increase in maintenance costs or a reduction in system efficiency. Recent advances in machine learning (ML) techniques have encouraged condition monitoring researchers to focus their efforts on finding suitable ML-based solutions for system condition assessment. However, to address the reliability issue, it is crucial to consider a larger amount of data measured by heterogeneous sensors on the system together with non-sensor information. The trend of data fusion has already started in other areas of ML application, and many of today's state-of-the-art models benefit from various types of fusion techniques to improve their accuracy. However, traditional classifiers do not provide any information about the prediction uncertainty, and they tend to show falsely high confidence when encountering low-quality data or previously unseen classes. Fusion of different data sources without considering the epistemic or aleatory uncertainty can lead to a deterioration of the result. Bayesian frameworks have traditionally been used to quantify uncertainty of systems; however, only recent advances made it possible to successfully implement Bayesian ML models.

The research methodology was investigated using the MAFAULDA dataset generated by SpectraQuest's Machinery Fault Simulator. This simulator experimentally simulated various bearing conditions, including normal operation and inner and outer ring bearing failures, at variable speeds. The dataset consists of 1951 instances measured using two triaxial accelerometers, a microphone, and a tachometer.

Diagnosis has been done via two multi label 1D Convolutional Neural Networks - each for a selected sensor - and their prediction along with their associated uncertainty quantity has been fused utilizing Bayesian model averaging. The methodology is capable of fusion of various decisions made based on different data sources and generate a unified decision with associated confidence level. Fusion process is uncertainty aware and application of 1D networks reduce the amount of data needed.

1. INTRODUCTION

1.2. Motivation behind the study

While Condition Monitoring Systems (CMS) have been extensively researched in recent years, the issue of their reliability has often been overlooked. It's crucial to recognize that CM relies on a complex system comprising sensors,

acquisition devices, data analysis techniques, and expertise. Having in mind that any system that can fail, would eventually fail highlights the importance of reliability studies on CMS. This gap in research domain has prompted the authors to investigate the reliability of CMS with the aim of increasing awareness and attracting other scientists' attention to the issue of the reliability of CMS.

The main goal of condition monitoring is to enhance the detection of failures compared to traditional methods like periodic maintenance in a cost-effective manner [1]. Therefore, addressing uncertainty in models does not imply admitting their malfunction; rather, a CMS with high uncertainty may prevent severe failures and associated costs and casualties that could be overlooked by competitor approaches. Acknowledging and addressing sources and levels of uncertainty in any diagnosis system is essential, as uncertainty is an inevitable aspect. Providing this critical information can help operators to make informed decisions and conduct thorough risk analyses.

1.1. Condition monitoring background

Condition monitoring (CM) serves as a vigilant process or a precision instrument focused on the early detection of machinery faults, failures, and wear, aiming to minimize downtimes and maintenance costs while maximizing production output. By detecting failures in their early stages, CM optimizes maintenance planning and action, thereby mitigating the risk of escalating damage and catastrophic failures. Moreover, it enhances comprehension of machinery behavior, consequently refining maintenance practices and operational efficiency [2].

CM techniques typically involve continuous measurement of machinery indicators or signals (online CM) or periodic assessments at predetermined intervals (offline CM) to detect abnormal deviations from baseline signals, distinguishing them from normal operational variations or detecting any fault signature [3].

Many examples for the development of CM can be found in literature: [4] designed and developed an integrated wireless vibration sensing tool to monitor milling equipment, employing Support Vector Machine (SVM) for analysis. [5] compared the statistical parameters of vibration signals for bearing diagnosis and suggested that signal power is the most effective criterion for diagnosis. [6] proposed an intelligent feature extraction method from vibration signals of bearing datasets to prevent human intervention for large signal analysis tasks. [7] reviewed vibration based condition monitoring of rotary machinery. [8] proposed a deep learning based gearbox fault diagnosis method that addresses data scarcity. [9] have fused multiple vibration signal into two-dimensional rectangular matrix and employed a two-dimensional convolutional neural network (2D-CNN) for bearing fault diagnosis.

Rotating machinery is a primary focus of CM research due to its challenging nature. This includes various industrial components such as rolling and journal bearings, gearboxes, shafts, blades, entire systems like wind turbines, reciprocating machines, electric motors, pumps, helicopters, fans, cam mechanisms, generators, and compressors. Various diagnostic parameters can be monitored, such as vibrations, acoustic emissions, electrical currents, flow rates, rotational speeds, pressure levels, temperature, lubrication conditions, strain, wear, and rotor-stator interactions. Vibration emerges as the predominant condition indicative of rotary machine health, as each component exhibits a unique vibration signature closely correlated with operational conditions. Faults or defects within components introduce additional dynamic forces, manifesting as vibrations within specific frequency ranges. Notable fault types detectable via vibration-based CM techniques include looseness, eccentricity, unbalance, blade defects, misalignment, bearing faults, gear damage, and shaft deformations [3].

1.2. Uncertainty in diagnosis

Uncertainty plays a significant role in human affairs, permeating everyday decisions in ordinary life. Decision-making, a fundamental capability of human beings, is essential for survival and well-being. However, decision-making is inherently challenged by uncertainty about the future. Anticipation of future events, upon which decisions are based, is inevitably subject to uncertainty. This is particularly evident in diagnostic uncertainty in engineering, where engineers often struggle to make definitive diagnoses despite extensive testing and relevant information [10].

In the realm of CM, ensuring the reliability of the diagnostic system is paramount. Indicating a fault where no fault is (a so-called false positive) can lead to unnecessary stoppages and maintenance, increasing operational costs, while false negatives risk failure and the propagation of damage. Ensuring the reliability of CMS is crucial for achieving their main goals of cost reduction and failure prevention. Considering the substantial investment necessary for implementing these techniques, only a reliable system that effectively prevents expensive failures can be justified.

To address these challenges, an uncertainty-aware fusion approach is essential. This approach involves explicitly modeling and quantifying the uncertainty associated with each source of information and the fusion process itself. By accounting for uncertainty, decision-makers can better assess the reliability and confidence level of the fused information. Moreover, an uncertainty-aware fusion approach enables the identification of potential sources of error or bias in the fusion

process, allowing for more robust and trustworthy decision-making outcomes.

Information fusion, as a methodological approach, presents a promising solution to the challenge of managing uncertainty in complex systems. By integrating diverse sources of information, including both sensory and non-sensory data, information fusion aims to enhance decision-making processes by providing a more comprehensive and accurate understanding of the observed system [11].

One of the primary advantages of information fusion is its capacity to leverage the strengths of individual sources of information while compensating for their inherent limitations. For instance, while sensory data such as vibration measurements may provide insights into the mechanical condition of a machine, non-sensory data such as operational logs or historical maintenance records can offer valuable contextual information. By combining different types of information, information fusion enables a more holistic assessment of the system's health status. However, implementing data fusion poses several challenges, particularly due to the diversity of data sources and sensor technologies involved. These challenges include issues related to data compatibility, data quality, and data integration. For instance, data collected from different sensors may vary in terms of accuracy, precision, and sampling frequency, making it challenging to effectively merge them into a cohesive dataset. Neglecting model uncertainty during fusion process can significantly impact the reliability of the fused information. Inaccurate or unreliable diagnoses from individual sources can propagate errors and inconsistencies throughout the fusion process, leading to a loss of fidelity in the final fused output. [12]

1.3. Authors contribution

The field of condition monitoring is vast, with numerous research initiatives aiming to enhance fault diagnosis techniques. This work contributes to the existing body of knowledge by introducing several approaches:

1- Multi-Label Fault Diagnosis: The authors propose a multi-label approach to fault diagnosis, enabling handling of complex fault scenarios. This methodology allows for the assignment of independent probability values to each fault class, providing a more detailed understanding of the system's health status.
2- Addressing Data Scarcity: The research addresses the common challenge of data scarcity by introducing a Custom 1D Convolutional Neural Network (CNN). 1D CNN architecture reduces the amount of data required for accurate fault diagnosis, thereby overcoming limitations associated with insufficient data availability.
3- Reliability Enhancement: The study enhances the reliability of fault diagnosis by leveraging multiple probabilistic decisions from different sensors. Through a Bayesian Model Averaging (BMA) approach, the authors combine the probabilistic outputs of various sensors, resulting in more robust and accurate diagnostic outcomes. This integration of diverse sensor data contributes to improved decision-making and system health assessment.

2. MULTILABEL PROBLEM

In traditional single-label classification, the model learns from a set of examples, each associated with a single label from a set of distinct labels. Typically, a traditional classifier utilizing a SoftMax layer assigns a probability value to each label, ensuring that the sum of probabilities across all possible labels equals one. The model then selects the label with the highest probability as the predicted label. However, this approach limits the model to predicting only one label per instance.

In contrast, in multi-label classification, the model assigns a probability value between zero and one independently to each class for a given instance. This allows for multiple labels to simultaneously have high probabilities. These classes are non-mutually exclusive and may overlap by definition. This approach was mainly used for text categorization and medical diagnosis [13]. Multi-label classification has been used by [14] to classify X-ray image via residual attention learning to diagnosis thorax disease. [15] utilized multi-label modeling for person re-identification to address the challenges of unsupervised learning, utilizing memory-based non-parametric classifier and integrates multi-label classification and single-label classification in a unified framework. [16] used attention based multi-label graph neural network to highlight the dependencies of labels in text classification.

In the context of system diagnosis and machinery fault detection, multi-label classification has not been widely investigated despite its value for identifying complex faults, especially when there's a correlation between them. By setting an appropriate threshold, the model can predict a neutral class, indicating uncertainty about the outcome, rather than forcing a specific label prediction. In contrast to single-label classification where the model assigns a probability value summing up to one across all classes, multi-label classification assigns an independent probability to each label (see Figure 1). This means that there may be cases where none of the labels have a high enough probability to cross the

threshold, indicating that the network lacks confidence in its prediction.

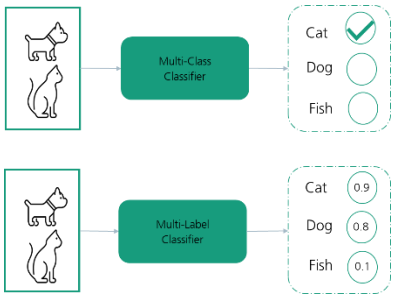


Figure 1 Label assignment in Multi-class vs Multi-label modeling

Various techniques exist for implementing a multi-label model, as shown in Figure 2. However, the details of each technique are beyond the scope of this paper, and interested readers are referred to [13, 17] for more information.

In addressing the problem, two primary approaches are commonly employed. Firstly, we can transform the problem into smaller, single-label components, allowing the use of any machine learning method to address each segment. Alternatively, we can adapt the algorithm itself to enable multi-label classification.

In the transformation approach, we can convert the problem into single-label binary classification using various methods:

- Powerset of Labels: This method decomposes the problem into all possible combinations of labels. While it provides insight into label relations, it can be computationally expensive.
- Binary Relevance: This approach compares a single label to all others or to one other label.
- Label Manipulation: We can also delete or create new labels as needed.

When implementing CNNs, different loss functions and activation layers may be required at the end of the network to accommodate multi-label classification.

3. BAYESIAN MODEL AVERAGING

In many cases, multiple models can adequately describe the distributions that generate observed data. When faced with this scenario, selecting the best model becomes crucial and is typically based on criteria such as how well the model fits the observed dataset, its predictive capabilities, or likelihood penalizations like information criteria. Once a model is selected, inferences are drawn and conclusions are made under the assumption that the selected model accurately represents the underlying truth. However, there are drawbacks to this approach. Selecting a single model can lead to overconfident inferences and riskier decisions, as it overlooks the inherent uncertainty in model selection and relies heavily on specific assumptions about the selected model. [18]

BMA provides a systematic and coherent methodology for addressing model uncertainty. It applies Bayesian inference directly to the problem of model selection, combined estimation, and prediction. BMA provides a straightforward criterion for model selection and leads to more cautious predictions. However, implementing BMA can be challenging, as it involves making various assumptions and decisions based on specific situations and contexts. [18]

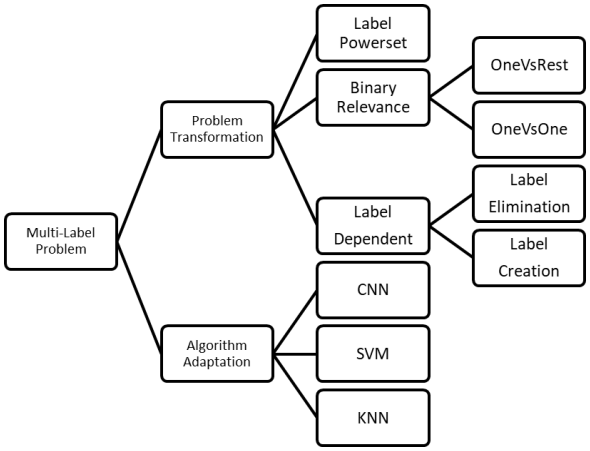


Figure 2 Overview of multi-label classification techniques

Let us consider an ensemble of models represented as $\boldsymbol{M_l}, \boldsymbol{l} = \boldsymbol{1}, \dots, \boldsymbol{K}$, and let $\boldsymbol{Y}$ represent observed data from dataset and $\boldsymbol{\theta_l}$ be parameter of the model $\boldsymbol{l}$, then likelihood function of $\boldsymbol{Y}$ given $\boldsymbol{\theta_l}$ and $\boldsymbol{M_l}$ can be written as $\boldsymbol{L(Y|\theta_l, M_l)}$. Additionally, prior probability of model parameters neglecting hyperparameters can be written as $\boldsymbol{\pi(\theta_l|M_l)}$ now, one can easily show posterior probability for model parameters as:

$$\pi(\theta_l|Y, M_l) = \frac{L(Y|\theta_l, M_l)\pi(\theta_l|M_l)}{\int L(Y|\theta_l, M_l)\pi(\theta_l|M_l)\, d\theta_l} \quad (1)$$

The denominator of (1) is called model's marginal likelihood or model evidence which represent prior distribution of all the parameter values related to model $\boldsymbol{M_l}$. Let's denote it as:

$$\pi(Y|M_l) = \int L(Y|\theta_l, M_l)\pi(\theta_l|M_l)\, d\theta_l \quad (2)$$

Bayesian model averaging introduces an additional level to this hierarchical modeling framework by incorporating a prior distribution over the entire set of models under consideration. This incorporates for the prior uncertainty regarding each model's ability to accurately represent the observed data. This is represented as a probability density function across all the models, and can be written as $\boldsymbol{\pi(M_l)}$

or $\boldsymbol{l = 1, \ldots, K}$, now we can show the posterior of model probability as:

$$\pi(M_l|Y) = \frac{\pi(Y|M_l)\pi(M_l)}{\Sigma_{l=1}^{k}\pi(Y|M_l)\pi(M_l)} \quad (3)$$

One now may re-write (3) as a ratio to a baseline model:

$$BF_{lm} = \frac{\pi(M_l|Y)}{\pi(M_m|Y)} \quad (4)$$

This can be interpreted as the relative strength of the models with respect to each other. It is clear that Eq. (3) can be expressed as the division of Eq. (4) as: [18]

$$\pi(M_l|Y) = \frac{BF_{l1}\pi(M_l)}{\Sigma_{m=1}^{k}BF_{m1}\pi(M_m)} \quad (5)$$

If $\boldsymbol{\Delta}$ is a quantity of interest, such as the utility of a course of action, then its posterior distribution can be formulated as: [19]

$$\pi(\Delta|Y) = \Sigma_{l=1}^{k}\pi(\Delta|M_l, Y)\pi(M_l|Y) \quad (6)$$

Here and on for simplicity we would address $\boldsymbol{\pi(M_l|Y)}$ term as $\boldsymbol{w_l}$. The $\boldsymbol{w_l}$s are probabilities; hence, they are nonnegative and sum up to 1. It is important to bear this in mind during their estimation. [20]

4. ESTIMATING BY LIKELIHOOD MAXIMIZATION

For convenience, we restrict attention to the situation where the conditional probability density functions (PDFs) are approximated by normal distributions. We maximize $\boldsymbol{w_k}$ by maximum likelihood from the validation/training dataset. The likelihood function is defined as the probability of the training data given the parameters to be estimated. The maximum likelihood estimator is the value of the parameter vector that maximizes the likelihood function, that is, the value of the parameter vector under which the observed data were most likely to have been observed. It is convenient to maximize the logarithm of the likelihood function (or log-likelihood function) rather than the likelihood function itself, for reasons of both algebraic simplicity and numerical stability; the same parameter value that maximizes one also maximizes the other. Estimation through likelihood maximization involves approximating the conditional PDFs here for ease of computation normal distributions has been selected. We maximize the weights $\boldsymbol{w_k}$ by maximizing the likelihood function using the validation/training dataset. The likelihood function represents the probability of observing the training data given the parameters to be estimated.

$$L(w_k|Y) = \Sigma_t \Sigma_{k=1}^{k}\, log\, \pi(\Delta|M_l, Y)\, w_k \quad (7)$$

where the summation is over values of $\boldsymbol{t}$ that index observations in the training set. [20]

5. MODEL ARCHITECTURE

5.1. Convolutional neural network

CNNs have received considerable attention and have been proven effective in various domains. One promising area for CNNs is in fault diagnosis and CM. Researchers have been increasingly using ML techniques, especially CNNs, for system diagnosis, particularly when monitoring signals such as vibration, acoustics, or temperature. [21] has utilized multi-branch residual convolutional neural network to diagnose crane gearbox with vibration signal that has been transferred to 2D images using Markov transformation field. [22] suggested an explainable CNN model that analysis cyclostationary vibration signals to diagnose wind turbine gearbox fault. [23] has proposed a light weight CNN model for bearing fault diagnosis based on Fast Fourier Transfer

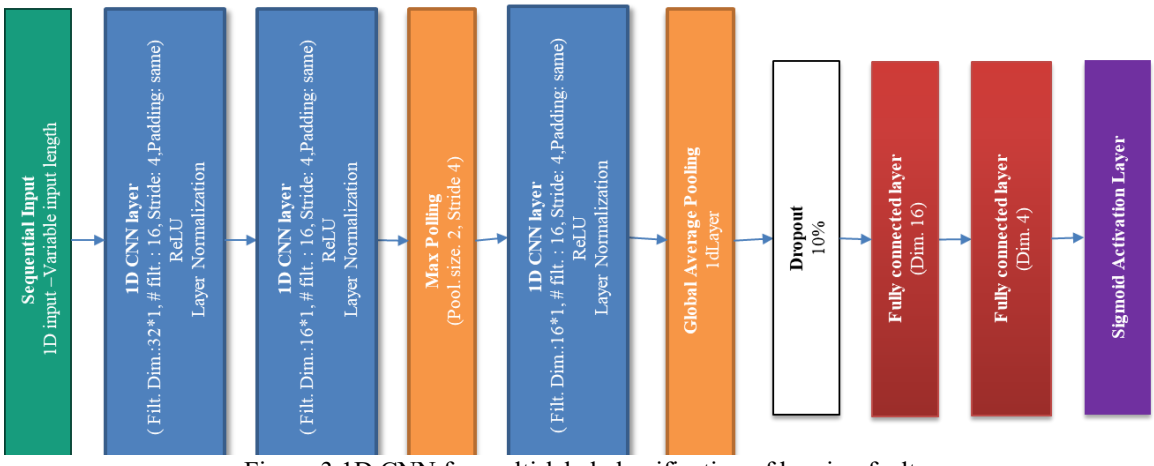


Figure 3 1D CNN for multi-label classification of bearing fault

(FFT) image coding of vibration signals. [24] has proposed a multiscale quadratic attention-embedded CNN with attention mechanisms to address the challenges associated with bearing vibration signals for fault diagnosis. [25] has fused vibration and microphone signals utilizing a 1D-CNN to enhance the accuracy of diagnosis. [26] has introduced a CNN model to diagnose bearing fault utilizing motor speed signal to remove the necessity of additional sensors.

A CNN consists of several layers, including an input layer, a convolutional layer, an activation layer, and a fully connected layer. Additional layers such as normalization and dropout are often used for generalization and to prevent overfitting. At the core of CNNs are convolutional layers, which allow us to automatically extract features from input data by mimicking how the brain's visual cortex processes images. This can be achieved by convoluting the input data with a filter, which is an n by m matrix whose elements are defined during the training phase, and moving the filter through the data at a constant step called a "stride". The convolution layer produces new images called feature maps. The feature map emphasizes the unique features of the original image. [27, 28] Although 2D CNNs have been commonly used for vibrational based diagnosis tasks, their effectiveness depends on a preprocessing step that converts the 1D signal into a 2D format. However, this preprocessing step often results in information loss and reduced diagnostic reliability. Although 1D and 2D CNNs share similar architectures, the key difference between them lies in their filter sliding mechanisms. In 1D CNNs, the filter slides vertically along the height to extract features, with the height determining the number of sample points for convolutional operations. On the other hand, 2D CNNs slide the filter both horizontally and vertically, with the height and width of the filter dictating the range of convolution operations for each step. However, 1D CNNs offer advantages over their 2D counterparts when processing 1D signals. This preference stems from several factors: [29]

- Computational complexity of 1D and 2D convolution calculations differ due to the fact that 1D CNN operates with one dimension less, resulting in significantly lower computational costs under identical conditions (same configuration, network, and hyperparameters).
- Reduced computational complexity makes 1D CNN suitable for low-cost real-time applications on smaller devices.
- Processing signals in the time domain eliminates the need for an additional step to convert a one-dimensional signal to a two-dimensional signal. This avoids adding irrelevant data and preserves the information present in the original data.

Here, we introduce a customized 1D CNN network (refer to Figure 3) along with the associated hyperparameters (see Table 1) for multi-label classification of bearing fault diagnosis. The application of 1D CNN allows us to employ shallower networks and avoids the inclusion of irrelevant information that may result from the conversion of 1D to 2D data.

By employing a sigmoid activation function at the last layer of the CNN architecture, along with a binary entropy loss function, the conventional multi-class CNN classifier has been transformed into a multi-label classifier that operates independently within each class and predicts whether the instance belongs to that class or not, as in a "one against all" strategy. This approach eliminates the need to train multiple networks for each label, thus reducing the necessity of large data and computational effort.

Table 1 Model Hyperparameters

Hyperparameter	Value
Mini batch size	25
Max epoch	50
Network selection (Early stoppage)	Minimum validation loss
optimizer	Adam
Learning rate	0.001
Loss Function	Binary cross-entropy
Padding	“Same”
Software	MATLAB

6. TEST DATASET AND PREPRATION

The methodology has been applied on the MAFAULDA dataset. The dataset consists of 1951 multivariate time-series acquired by sensors on SpectraQuest's Machinery Fault Simulator (MFS) Alignment-Balance-Vibration (ABVT). It includes six different simulated states: normal function, imbalance fault, horizontal and vertical misalignment faults, and inner and outer bearing faults. This heterogeneous dataset involves measuring acoustic and vibration signals, providing comprehensive insights into machinery behavior and fault diagnosis. Each measurement lasts for 5 seconds, with 49 measurements for normal conditions, 197 for horizontal misalignment with angles of 0.5, 1.0, 1.5, and 2.0 degrees, 301 for vertical misalignment with angles of 0.51, 0.63, 1.27, 1.40, 1.78, and 1.90 degrees, and 333 for mass imbalance of 6, 10, 15, 20, 25, 30, and 35 grams. Bearing faults have been

combined with 5, 6, 20, and 35 grams of mass imbalance to enhance the effect of the fault. The available experiment specification includes details of used equipment's, including the SpectraQuest Inc. Alignment/Balance Vibration Trainer (ABVT) Machinery Fault Simulator (MFS), Industrial IMI Sensors accelerometers, Monarch Instrument MT-190 analog tachometer, and Shure SM81 microphone. Data acquisition parameters such as sensitivity, frequency range, and measurement range are specified for each sensor. Sequences are categorized based on fault types, with details on the number of sequences per fault category, load values, and degrees of misalignment. The database is openly accessible online, with links provided at [30] for downloading the entire dataset or specific parts corresponding to different fault types. Figure 4 depicts the data preparation process for training the models. Raw vibration signals from the tangential direction of the overhang (sensor number four) and underhang (sensor number seven) accelerometers, each corresponding to a different model, are inputted along with the tachometer signal. These signals are then divided into five successive parts of one second each. The first three rotations of each one-second signal are then extracted, resulting in variable vector lengths. Following, random noise is added to reduce signal quality to signal-to-noise ratio (SNR) level of 10. The data set is then randomly divided into training (60 %), validation (20 %), and test (20 %) sets to facilitate model evaluation and validation. Additionally, reducing the data to three revolutions per second helps to evaluate the model under more realistic conditions where acquiring large datasets may not be feasible.

7. RESULT

The proposed 1D CNN was trained using the preprocessed training set (see Figure 4) of data from sensors four and seven separately. Probability acceptance threshold of 0.5 was set for each label output by the models. The models were then evaluated on the test dataset, and the performance results for the tangential overhang accelerometer signal are reported in Table 2, while those for the tangential underhang accelerometer are shown in Table 3. The corresponding confusion matrices are depicted in Figure 5 and Figure 6. Subsequently, BMA was performed on the two models, where BMA parameters were computed by maximizing the likelihood using the validation dataset. The related values are reported in Table 4. Finally, the results of the combined model via BMA, with the same probability acceptance threshold of 0.5, are shown in Table 5, along with its confusion matrix in Figure 7. Two instances from the test set have been selected and reported in Table 6 to demonstrate the step-by-step improvement of the results:

- In case A, the underhung model shows the highest probability for the outer race fault, which is an incorrect label. However, overhung and the combined model correctly identifies the fault as a cage fault.
- Case B reports an instance where the underhung model correctly identifies the label, but the overhung model fails to do so. Once again, the combined model correctly classifies the instance.

The results indicate an increase in performance of almost 5 % over the overhang accelerometer model and an increase of 8 % over the underhang accelerometer model. Considering the confusion matrix and accuracy of each class for each model, the calculated BMA parameters were as expected.

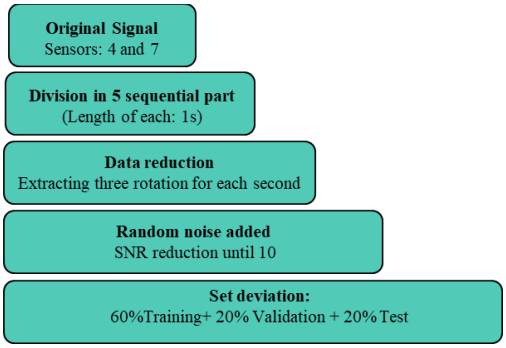


Figure 4 Data preparation scheme

Table 2 Accuracy of proposed 1D multi-label CNN for tangential Overhang accelerometer

Tangential Overhang accelerometer				
Label	Outer Race Fault	Cage Fault	Ball Fault	Healthy
Accuracy (%)	78.06	97.45	92.35	99.49
Overall Accuracy (%)			87.76	

Table 3 Accuracy of proposed 1D multi-label CNN for sensor tangential Underhung accelerometer

Tangential Underhung accelerometer				
Label	Outer Race Fault	Cage Fault	Ball Fault	Healthy
Accuracy (%)	95.41	89.29	95.41	99.49
Overall Accuracy (%)			84.69	

Table 4 BMA parameters

Posterior probability of Overhang Model	Posterior probability of Underhung Model
0.3728	0.6272

Table 5 Accuracy of combined model via BMA

Tangential Underhung accelerometer				
Label	Outer Race Fault	Cage Fault	Ball Fault	Healthy
Accuracy (%)	93.88	96.43	96.94	99.49
Overall Accuracy (%)			91.84	

Table 6 Instances from test set

	Case A			Case B		
	Overhang	Underhung	Combined	Overhang	Underhung	Combined
Outer Race Fault	0.04	0.59	0.39	0.46	0.78	0.66
Cage Fault	1.00	0.24	0.52	0.01	0.2	0.13
Ball Fault	0.00	0.05	0.03	0.54	0.02	0.21
Healthy	0.00	0.00	0.00	0.00	0.00	0.00
True label	Cage Fault			Outer Race Fault		

8. CONCLUSION

This study introduces a multi-label approach to fault diagnosis, which facilitate the handling of complex fault scenarios by assigning an independent probability value to each class. To address the common issue of data scarcity, a Custom 1D CNN is proposed to reduce the required amount of data. Additionally, a BMA approach is employed to enhance the reliability of diagnosis by combining multiple decisions from different sensors. Evaluation of the technique on a public dataset shows a 5 to 8 % improvement in the accuracy of the combined BMA model result compared to individual models. The discussed algorithm provides an explainable process for decision fusion, emphasizing the quality of each diagnosis. BMA offers an uncertainty-aware fusion platform, where each model contributes based on its performance in the training and validation phases.

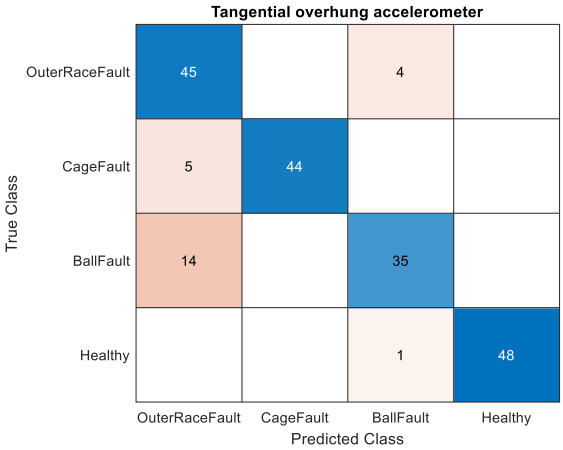


Figure 5 Confusion Matrix Multi-Label classifier - Tangential Overhung accelerometer

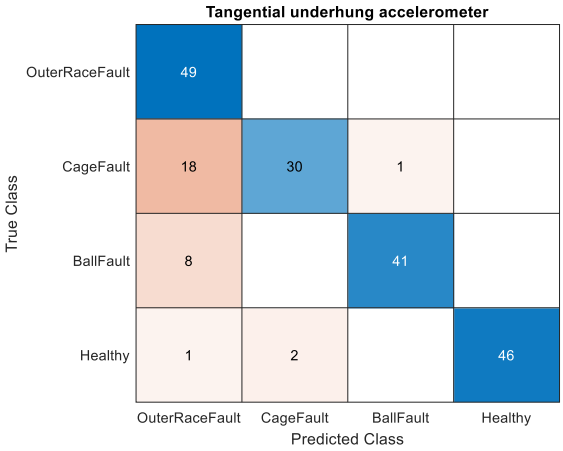


Figure 6 Confusion Matrix Multi-Label classifier - Tangential Underhung accelerometer

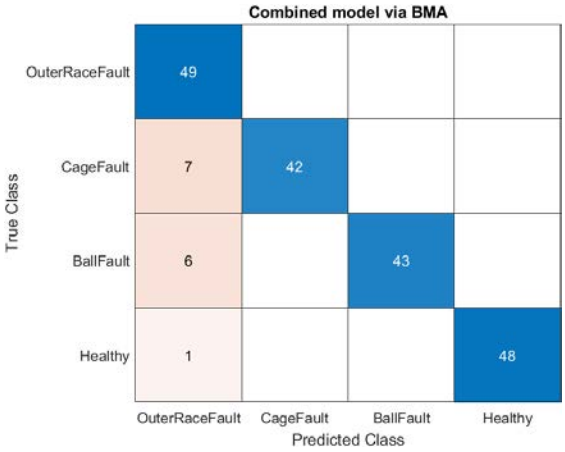


Figure 7 Confusion Matrix BMA combined Model

ACKNOWLEDGEMENT

The Authors Gratefully acknowledge the European commission for its support of the Marie Sklodowska Curie program through the ETN MOIRA project (GA 955681).

References

[1] David McMillan and Graham Ault, Eds., *Towards Quantification of Condition Monitoring Benefit for Wind Turbine Generators*, 2007.

[2] H. D. M. de Azevedo, A. M. Araújo, and N. Bouchonneau, "A review of wind turbine bearing condition monitoring: State of the art and challenges," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 368–379, 2016, doi: 10.1016/j.rser.2015.11.032.

[3] M. Tiboni, C. Remino, R. Bussola, and C. Amici, "A Review on Vibration-Based Condition Monitoring of Rotating Machinery," *Applied Sciences*, vol. 12, no. 3, p. 972, 2022, doi: 10.3390/app12030972.

[4] C. Zhou, K. Guo, and J. Sun, "An integrated wireless vibration sensing tool holder for milling tool condition monitoring with singularity analysis," *Measurement*, vol. 174, p. 109038, 2021, doi: 10.1016/j.measurement.2021.109038.

[5] N. Tandon, "A comparison of some vibration parameters for the condition monitoring of rolling element bearings," *Measurement*, vol. 12, no. 3, pp. 285–289, 1994, doi: 10.1016/0263-2241(94)90033-7.

[6] F. Jia *et al.,* "A method of automatic feature extraction from massive vibration signals of machines," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 2016, pp. 1–6.

[7] C. Mongia, D. Goyal, and S. Sehgal, "Vibration response-based condition monitoring and fault diagnosis of rotary machinery," *Materials Today: Proceedings*, vol. 50, pp. 679–683, 2022, doi: 10.1016/j.matpr.2021.04.395.

[8] S. R. Saufi, Z. A. B. Ahmad, M. S. Leong, and M. H. Lim, "Gearbox Fault Diagnosis Using a Deep Learning Model With Limited Data Sample," *IEEE Trans. Ind. Inf.*, vol. 16, no. 10, pp. 6263–6271, 2020, doi: 10.1109/TII.2020.2967822.

[9] J. Wang, D. Wang, S. Wang, W. Li, and K. Song, "Fault Diagnosis of Bearings Based on Multi-Sensor Information Fusion and 2D Convolutional Neural Network," *IEEE access : practical innovations, open solutions*, vol. 9, pp. 23717–23725, 2021, doi: 10.1109/ACCESS.2021.3056767.

[10] G. J. Klir, *Uncertainty and information: Foundations of generalized information theory*. Hoboken N.J.: Wiley-Interscience, 2006.

[11] S. Hassani, U. Dackermann, M. Mousavi, and J. Li, "A systematic review of data fusion techniques for optimized structural health monitoring," *Information Fusion*, vol. 103, p. 102136, 2024, doi: 10.1016/j.inffus.2023.102136.

[12] D. Wheeler, Esther D Meenken, Martin Espig, and Mos Sharifi, Eds., *UNCERTAINTY -WHAT IS IT?*, 2020.

[13] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas, "A Review of Multi-Label Classification Methods,"

[14] Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, pp. 259–266, 2020, doi: 10.1016/j.patrec.2018.10.027.

[15] D. Wang and S. Zhang, "Unsupervised Person Re-identification via Multi-label Classification," 20-Apr-20. [Online]. Available: http://arxiv.org/pdf/2004.09228

[16] A. Pal, M. Selvakumar, and M. Sankarasubbu, "MAGNET: Multi-Label Text Classification using Attention-based Graph Neural Network," *12th International Conference on Agents and Artificial Intelligence (ICAART*, pp. 494–505, 22-Mar-2020, doi: 10.5220/0008940304940505.

[17] A. C. P. L. F. de Carvalho and A. A. Freitas, "A Tutorial on Multi-label Classification Techniques," in *Studies in Computational Intelligence*, v. 201-206, *Foundations of computational intelligence*, A. E. Hassanien, Ed., Berlin: Springer, 2009-, pp. 177–195.

[18] T. M. Fragoso, W. Bertoli, and F. Louzada, "Bayesian Model Averaging: A Systematic Review and Conceptual Classification," *Int Statistical Rev*, vol. 86, no. 1, pp. 1–28, 2018, doi: 10.1111/insr.12243.

[19] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery and Chris T. Volinsky, "Bayesian Model Averaging: A Tutorial,"

[20] ADRIAN E. RAFTERY, TILMANN GNEITING, FADOUA BALABDAOUI, and MICHAEL POLAKOWSKI, "Using Bayesian Model Averaging to Calibrate Forecast Ensembles,"

[21] J. Zhang, Q. Zhang, X. Qin, and Y. Sun, "Robust fault diagnosis of quayside container crane gearbox based on 2D image representation in frequency domain and CNN," *Structural Health Monitoring*, vol. 23, no. 1, pp. 324–342, 2024, doi: 10.1177/14759217231168877.

[22] A. Amin, A. Bibo, M. Panyam, and P. Tallapragada, "Wind Turbine Gearbox Fault Diagnosis Using Cyclostationary Analysis and Interpretable CNN," *J. Vib. Eng. Technol.*, vol. 12, no. 2, pp. 1695–1705, 2024, doi: 10.1007/s42417-023-00937-1.

[23] K. Cui, M. Liu, and Y. Meng, "A new fault diagnosis of rolling bearing on FFT image coding and L-CNN," *Meas. Sci. Technol.*, vol. 35, no. 7, p. 76108, 2024, doi: 10.1088/1361-6501/ad3295.

[24] Y. Tang, C. Zhang, J. Wu, Y. Xie, W. Shen, and J. Wu, "Deep Learning-Based Bearing Fault Diagnosis Using a Trusted Multiscale Quadratic Attention-Embedded Convolutional Neural Network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024, doi: 10.1109/TIM.2024.3374311.

[25] G. Li, J. Wu, C. Deng, Z. Chen, and X. Shao, "Convolutional Neural Network-Based Bayesian Gaussian Mixture for Intelligent Fault Diagnosis of Rotating Machinery," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2021.3080402.

[26] Z. Guo, M. Yang, and X. Huang, "Bearing fault diagnosis based on speed signal and CNN model," *Energy Reports*, vol. 8, pp. 904–913, 2022, doi: 10.1016/j.egyr.2022.08.041.

[27] P. Kim, *MATLAB deep learning: With machine learning, neural networks and artificial intelligence*. New York, Berkeley, California: Springer; Apress, 2017. [Online]. Available: http://www.springer.com/de/book/9781484228456

[28] Neena Aloysius and Geetha M, *A Review on Deep Convolutional Neural Networks: 6th-8th April 2017, Melmaruvathur, India*. Piscataway, NJ, U.S.A.: IEEE, 2017. [Online]. Available: http://ieeexplore.ieee.org/servlet/opac?punumber=8269242

[29] R. F. R. Junior, I. A. d. S. Areias, M. M. Campos, C. E. Teixeira, L. E. B. Da Silva, and G. F. Gomes, "Fault detection and diagnosis in electric motors using 1d convolutional neural networks with multi-channel vibration signals," *Measurement*, vol. 190, p. 110759, 2022, doi: 10.1016/j.measurement.2022.110759.

[30] *MAFAULDA :: Machinery Fault Database [Online].* [Online]. Available: https://www02.smt.ufrj.br/~offshore/mfs/page_01.html (accessed: 06-Feb-24).

Prognosis of Internal Short Circuit Formation in Lithium-Ion Batteries: An Integrated Approach Using Extended Kalman Filter and Regression Model

Lorenzo Brancato[1], Yiqi Jia[2], Marco Giglio[3], and Francesco Cadini[4]

[1] *Politecnico di Milano, Department of Mechanical Engineering, Via La Masa 1, 20156, Milan, Italy*
lorenzo.brancato@polimi.it
yiqi.jia@polimi.it
marco.giglio@polimi.it
**Corresponding author: francesco.cadini@polimi.it*

ABSTRACT

The global transition to electric power, aimed at mitigating climate change and addressing fuel shortages, has led to a rising usage of lithium-ion batteries (LIBs) in different fields, notably transportation. Despite their many benefits, LIBs pose a critical safety concern due to the potential for thermal runaway (TR), often triggered by spontaneous internal short circuit (ISC) formation. While extensive research on LIB fault diagnosis and prognosis exists, forecasting ISC formation in batteries remains unexplored. This paper presents a new methodology that combines the extended Kalman filter (EKF) algorithm for real-time estimation of ISC state with an adaptive linear regressor model for forecasting remaining useful life (RUL). This approach is designed for seamless integration into actual battery management systems, offering a computationally efficient solution. Numerical validation of the framework was conducted due to the current lack of experimental data in the literature. The significance of this work lies in its contribution to ISC prognosis, providing a practical solution to enhance battery safety.

1. INTRODUCTION

In response to the increasing environmental consciousness and the urgent need to address climate change, car manufacturers and consumers are turning towards cleaner alternatives to traditional gasoline-powered vehicles. Electric vehicles (EVs) are at the forefront of this shift, offering significant reductions in emissions that lead to cleaner air and a more sustainable planet. This movement is not just a trend; governments worldwide are actively supporting and encouraging the adoption of EVs through various policies. These include the implementation of stricter emissions regulations, mandates for zero-emission vehicles, and substantial investments in charging infrastructure.

LIBs have emerged as the go-to choice power source in EVs due to their numerous advantages, such as high energy density, powerful performance, and extended lifespan (Ding, Cano, Yu, Lu, & Chen, 2019). Despite their many benefits, LIBs are subjected to degradation phenomena (Han et al., 2019). This continuous degradation poses risks like battery failures and safety hazards, above all TR accidents (Feng, Ouyang, et al., 2018). The most common cause of TR incidents is ISC, making it imperative for the battery management system (BMS) to detect ISC formation and prevent severe ISC formation early. This is pivotal for ensuring the safe and reliable operation of EVs.

Understanding the intricate mechanism behind spontaneous ISC formation is an ongoing area of study that demands further research (Feng, Ouyang, et al., 2018). However, observations indicate that ISC formation, when not triggered by external factors like crushing or penetration, generally progresses slowly (Zhang et al., 2021). Moreover, research has shown that ISC formation predominantly impacts the electrical and thermal properties of the cell (Huang et al., 2021). This suggests that monitoring both the voltage and temperature of the cell, which are typically available in commercial BMS, could be exploited to detect and track ISC formation.

In recent years, researchers have made significant strides in developing various diagnostic algorithms aimed at detecting ISC and preventing TR. Most of these approaches are purely data-driven and aim at identifying parameter inconsistencies among single or multiple cells. These approaches utilize factors such as voltage (Hermann & Kohn, 2013), temperature (Yang, Cui, & Wang, 2019), State of Charge (SoC) (Zheng et al., 2018), and capacity (Reichl & Hrzina, 2018). However,

the subtle changes in electro-thermal signals caused by spontaneous ISC formation may not be immediately discernible within the battery dynamic responses, especially during the early stages of ISC. Additionally, signal variations due to external factors could potentially trigger false alarms. Consequently, establishing a precise threshold value presents a considerable challenge, as this value greatly affects the speed and accuracy of detection.

More advanced data-driven approaches have emerged to address these challenges, harnessing the capabilities of machine learning techniques. These approaches employ models such as deep neural networks (Cui et al., 2024) or random forest (Liu, Hao, Han, Zhou, & Li, 2023). However, since these methods rely solely on existing data, their performance is significantly limited by the scarcity of available data and the difficulties in generating new datasets.

To overcome these limitations, also model-based approaches have been developed to detect ISC. These include equivalent circuit models (ECM) (Asakura, Nakashima, Nakatsuji, & Fujikawa, 2010; Yokotani, 2014; Ikeuchi, Majima, Nakano, & KASA, 2014; Feng, Pan, He, Wang, & Ouyang, 2018), or more advanced electro-chemical models (Ma, Deng, & Wang, 2023). The basic idea of these methods is to transform the problem of ISC detection into model parameters and state estimation. The battery models are established to predict the voltage and temperature of the cells. The measured voltage or temperature of each cell is then compared with the predicted value of the model. If the residual between the two exceeds the allowable error range, it is considered that an ISC has occurred.

This work fills a crucial gap in the literature by focusing on prognosis and predicting the behavior of batteries experiencing spontaneous ISC formation. While there are existing ISC detection methods, as far as the authors are aware, no other studies have delved into predicting the evolution of ISC. What sets apart the prognostic framework introduced in this work is its dual capability: not only does it detect ISC for early warnings, but it also quantifies its severity and forecasts its future progression, enabling timely preventive measures. Moreover, the methodology emphasizes efficiency in selecting models and algorithms, considering their practical implementation in a BMS.

We build the prognostic framework upon the capabilities of the online ISC estimation algorithm proposed by the same authors in Ref. (Jia, Brancato, Giglio, & Cadini, 2024), which, unlike other ISC detection methods:

- utilizes both electrical and thermal measurements to enhance ISC detection and estimation accuracy;
- detect ISC by estimating a model parameter strictly related to the spontaneous ISC formation, allowing also to track the ISC state evolution.

This study introduces a method for predicting the battery RUL probability density function (pdf) using an adaptive linear regressor model to forecast the evolution of the ISC state until an appropriate threshold is reached. The proposed method is designed to be fully automated and can be easily integrated into a BMS for diagnosing and prognosis of spontaneous ISC formation. Moreover, the flexibility of the framework lies in its capacity to accommodate, in principle, various ISC state trajectories.

To validate our approach, we conducted a numerical case study that simulated the effects observed in measurements due to spontaneous ISC formation. This study aims to evaluate the effectiveness of our framework, given the scarcity of experimental data. Gathering such data proves challenging due to the complex nature and associated risks inherent in this phenomenon.

The paper is structured as follows: Section 2 briefly details the methods employed in developing the diagnosis and prognosis framework. In Section 3, the capabilities of the proposed method are demonstrated through a numerical case study involving a cylindrical LIB cell experiencing spontaneous ISC formation. Finally, Section 4 presents the conclusions drawn from this work and suggests potential directions for further research.

2. METHODOLOGY

2.1. Online ISC estimation algorithm

A dynamical system state comprises variables describing its condition and behavior. Non-linear dynamical systems exhibit dynamics not expressible linearly. State-space models represent system dynamics and observations using a hidden state vector $\mathbf{x}$. State equation $\mathbf{f}$ governs state vector evolution with some input $\mathbf{u}$ and some process noise $\mathbf{w}$, while observation equation $\mathbf{h}$ relates observed data, denoted with $\mathbf{y}$, to the state vector $\mathbf{x}$, some input $\mathbf{u}$ and some measurement noise $\mathbf{n}$.

The EKF estimates non-linear system states (Simon, 2006). It approximates non-linear dynamics linearly via Taylor expansion. The algorithm involves two steps: prediction, estimating the next state ($\hat{\mathbf{x}}$) and observations ($\hat{\mathbf{y}}$) with the previous state and calculating error covariance matrix ($\Sigma_{\hat{\mathbf{x}}}$); correction, updating state estimate and covariance with weighted innovation terms based on system observations. Proper initialization of the algorithm is crucial, assigning values to state vector estimate and error covariance matrix. The full algorithm is detailed in Table 1.

The electro-thermal model of a cylindrical cell described in our previous work (Jia et al., 2024), whose governing equations and parameters are summarized in Table 2 and Table 3, is discretized in time considering the following augmented state vector that includes the equivalent ISC conductance parameter G_{ISC}, expressed in Ω^{-1}, and representative of the

Table 1. Description of the extended Kalman filter algorithm.

Extended Kalman Filter Algorithm
Initialization:
Initialize state estimate $\hat{\mathbf{x}}_0^-$ and error matrix covariance $\Sigma_{\hat{\mathbf{x}}_0^-}$
Prediction Step:
Predict the state estimate:
$\hat{\mathbf{x}}_k^- = \mathbf{f}(\hat{\mathbf{x}}_{k-1}^+, \mathbf{u}_{k-1}, \bar{\mathbf{w}}_{k-1})$
Predict the observations:
$\hat{\mathbf{y}}_k = \mathbf{h}(\hat{\mathbf{x}}_{k-1}^+, \hat{\mathbf{y}}_{k-1}, \mathbf{u}_{k-1}, \bar{\mathbf{n}}_{k-1})$
Predict the error covariance matrix:
$\Sigma_{\hat{\mathbf{x}}_k^-} = A_k \Sigma_{\hat{\mathbf{x}}_{k-1}^-} A_k^T + B_k \Sigma_{\mathbf{w}} B_k^T$
Correction Step:
Compute the Kalman gain matrix:
$K_k = \Sigma_{\hat{\mathbf{x}}_k^-} C_k^T (C_k \Sigma_{\hat{\mathbf{x}}_k^-} C_k^T + D_k \Sigma_{\mathbf{n}} D_k^T)^{-1}$
Update the state estimate:
$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + K_k(\mathbf{y}_k - \hat{\mathbf{y}}_k)$
Update the error covariance matrix:
$\Sigma_{\hat{\mathbf{x}}_k^+} = \Sigma_{\hat{\mathbf{x}}_k^-} - K_k(C_k \Sigma_{\hat{\mathbf{x}}_{k-1}^-} C_k^T + D_k \Sigma_{\mathbf{n}} D_k^T) K_k^T$
Where:
$A_k = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\big\|_{\hat{\mathbf{x}}_{k-1}^+, \mathbf{u}_{k-1}, \bar{\mathbf{w}}_{k-1}}$ $B_k = \frac{\partial \mathbf{f}}{\partial \mathbf{w}}\big\|_{\hat{\mathbf{x}}_{k-1}^+, \mathbf{u}_{k-1}, \bar{\mathbf{w}}_{k-1}}$
$C_k = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}\big\|_{\hat{\mathbf{x}}_{k-1}^+, \mathbf{u}_{k-1}, \bar{\mathbf{n}}_{k-1}}$ $D_k = \frac{\partial \mathbf{h}}{\partial \mathbf{n}}\big\|_{\hat{\mathbf{x}}_{k-1}^+, \mathbf{u}_{k-1}, \bar{\mathbf{n}}_{k-1}}$
and $\Sigma_{\mathbf{w}}, \Sigma_{\mathbf{n}}$ are the covariance matrices of the two independent, zero-mean, Gaussian processes **w** and **n**.

actual ISC state:

$$\mathbf{x} = [z,\ i_{RC_1},\ i_{RC_2}, h,\ G_{ISC},\ T]^T \tag{9}$$

where z is the dimensionless state-of-charge (SoC), i_{RC_1} and i_{RC_2} are the two polarization currents (expressed in A), and T is the surface temperature (expressed in K).

Finally, the input vector and the output vector are defined as follow:

$$\mathbf{u} = [i_t,\ v_t]^T \tag{10}$$

$$\mathbf{y} = [v_t,\ T]^T \tag{11}$$

assuming that the load current i_t, the terminal voltage v_t, and the surface temperature T are all measurable quantities.

2.2. RUL estimation via simple linear regression

A simple linear regression model describes the linear relationship between a dependent variable, y, also known as the response, and one independent variable, x, also known as the predictor. In general, a simple linear regression model can be a model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ i = 1, ..., n \tag{12}$$

where n is the number of observations, y_i is the i-th response, β_0 is the model constant, β_1 is the slope of the model, ε_i is the i-th error term that captures the variability in y_i that is not explained by the linear relationship with x_i (Seber & Lee, 2012). The usual assumptions for simple linear regression modeling are: (i) the error terms ε_i are uncorrelated; (ii) the error terms ε_i have independent and identical normal distributions with mean zero and constant variance, σ_ε^2; (iii) the responses y_i are uncorrelated.

In this study, we develop an approach to predicting the evolution of ISC state using a simple linear regression model. Here, the response variable is the estimated equivalent ISC resistance, denoted as R_{ISC}, which can be calculated as $1/G_{ISC}$. This estimate is obtained from the equivalent ISC conductance while cycling the battery cell. The predictor variable in our model is the number of cycles, denoted as N. To ensure adaptability, our approach involves fitting the simple linear regression model using a robust least squares estimation algorithm (Holland & Welsch, 1977) within a sliding window of fixed size W. This means that although the number of observations analyzed remains constant at $n = W$, the actual data points y_i can vary between each query. To address the uncertainties in our predictions, once the model is fit with the latest available data points, we generate different realizations of the ISC evolution by sampling from the estimated Gaussian distribution of the error terms, $\varepsilon_i \sim N(0, \sigma_\varepsilon)$. These realizations are then truncated when they reach a predetermined threshold for the ISC state value. Through this process, we estimate the pdf of the RUL.

3. RESULTS

In this section, we validate the performance of the framework proposed in this work through a numerical study due to the scarcity of experimental data in existing literature.

3.1. Simulating spontaneous ISC formation

To maintain simplicity and ensure consistency with our methodology, we use the electro-thermal battery model described in Section 2.1 to simulate the dynamics of a real battery cell. In practice, the voltage and surface temperature signals processed by the proposed online ISC estimation algorithm are generated by this same model, hereafter referred to as "the plant". Nonetheless, we equip the plant with appropriate noise generators to capture non-modeled dynamics.

In the plant, it is assumed that the progression of degradation follows a power-law pattern over time:

$$\begin{cases} R_{ISC}(t) = R_i - (R_f - R_i) \cdot \left(\frac{t}{T_{end}}\right)^{p(T)} \\ p(T) = p_0 \cdot e^{-\frac{c}{T}} \end{cases} \tag{13}$$

Table 2. The governing equations of the electro-thermal model.

Equation name	Equation expression	
Current Kirchhoff law	$i = i_t + i_{ISC}$, with $i_{ISC} = v_t / R_{ISC}$	(1)
Voltage Kirchhoff law	$v_t = \text{OCV}(z) + M_0 \text{sign}(i) + Mh - R_1 i_{R_1} - R_2 i_{R_2} - R_0 i$	(2)
Coulomb counting	$\frac{dz}{dt} = -\frac{\eta}{Q} i$	(3)
RC circuit dynamics	$\frac{di_{RC_j}}{dt} = -\frac{1}{R_j C_j} i_{RC_j} - \frac{1}{R_j C_j} i$, with $j = 1, 2$	(4)
Hysteresis dynamics	$\frac{dh}{dt} = -\left\lvert \frac{\gamma\eta}{Q} i \right\rvert h - \left\lvert \frac{\gamma\eta}{Q} i \right\rvert \text{sign}(i)$	(5)
Electrical heat	$Q_{in} = R_0 i^2 + \frac{v_t^2}{R_{ISC}}$	(6)
Dissipated heat	$Q_{out} = h_{conv}(T - T_{amb})A$	(7)
Heat balance	$\frac{dT}{dt} = -\frac{1}{mc_m}(Q_{in} - Q_{out})$	(8)

Where i_t is the load current, v_t is the terminal voltage, z is the state of charge, i_{RC_j} are the polarization currents, h is the unitless hysteresis state, and T is the surface temperature.

Table 3. Model parameters.

Parameter name	Symbol	Value	Unit
Open-circuit voltage	OCV	$f(z)$	V
Series resistance	R_0	9.18	mΩ
1st polarization resistance	R_1	2.53	mΩ
2nd polarization resistance	R_1	21.32	mΩ
1st polarization capacitance	C_1	5116	F
2nd polarization capacitance	C_1	3582	F
Instantaneous hysteresis voltage term	M_0	0	V
Dynamic hysteresis voltage term	M	0.16	V
Coulombic efficiency	η	0.994	–
Rate of decay constant	γ	1	–
Capacity	Q	2.05	Ah
Battery cell mass	m	76	g
Specific heat capacity	c_m	1095	J/KgK
Heat transfer convection coefficient	h_{conv}	10	$\text{W/m}^2\text{K}$
Battery outer surface	A	5.3×10^{-3}	m^2
Ambient temperature	T_{amb}	298	K

Here the exponent p is a variable that changes with temperature, behaving according to an Arrhenius function; R_i is the initial ISC resistance value; R_f is the final ISC resistance value; T_{end} is the cycling time needed to evolve from R_i to R_f; p_0 and c are respectively the Arrhenius constant and rate terms. The values of these parameters are indicated in Table 4

Table 4. Degradation model parameters.

Parameter name	Symbol	Value	Unit
Initial ISC resistance value	R_i	1000	Ω
Final ISC resistance value	R_f	0.1	Ω
Cycling time from R_i to R_f	T_{end}	1200	h
Arrhenius constant term	p_0	3278	–
Arrhenius rate term	c	3.1×10^{-3}	K

The degradation model, which also incorporates a temperature-dependent exponent, is formulated and parameterized based on the following assumptions:

- ISC persists throughout the entire lifespan of the cell, with its formation and evolution spanning hundreds of hours or more (Zhang et al., 2021);
- As ISC progresses, the internal temperature of the cell increases, leading to complex chemical reactions involving electrode materials, electrolyte, and separator (Feng, Ouyang, et al., 2018);
- Most of these chemical reactions are exothermic, accelerating the TR occurrence. When the temperature overcomes a critical point, various degradation phenomena occur, such as solid-electrolyte interphase layer decomposition, anode-electrolyte reactions, electrolyte breakdown, separator meltdown, and cathode failure. All these phenomena contribute to further increasing the internal temperature of the cell, ultimately triggering TR (Feng, Ouyang, et al., 2018).

The way the degradation sub-model described by Eq. (13) relates to the electro-thermal cell model summarized by the equations in Table 2 is graphically illustrated in Figure 1. This model is simulated by cycling the plant using a dynamic stress test current cycle and constant charging. The resulting ISC state evolution is illustrated at the top of Figure 2, where the

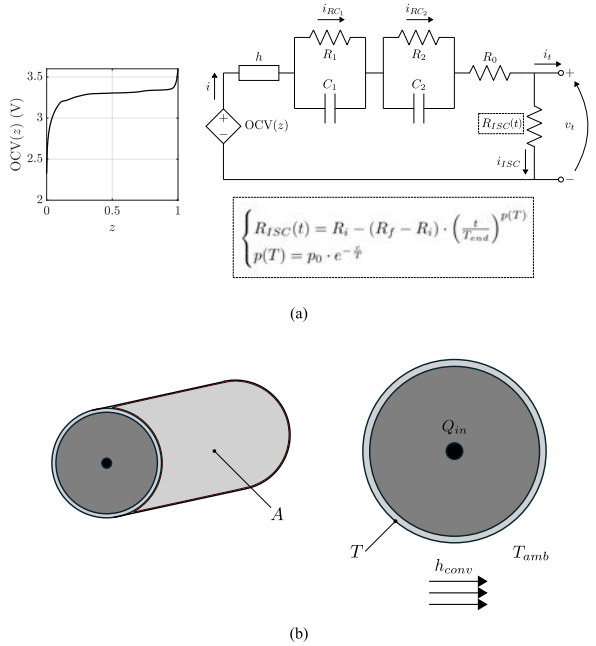


Figure 1. Electro-thermal battery cell model subjected to spontaneous ISC degradation model evolution. (a) Electrical sub-model coupled with the ISC degradation model. (b) Thermal sub-model.

x-axis has been scaled to cycles and the y-axis has been log-scaled to enhance visibility.

During the simulation, some important health-related quantities, i.e., the maximum temperature, T_{max}, and the discharging time, t_{dis}, measured across one cycle, have also been recorded along with the voltage and temperature measurements. These quantities are shown at the bottom of Figure 2. These quantities are, in fact, strongly correlated with the severity of the ISC. The observed trends agree with those expected for spontaneous ISC formation, as referenced in (Feng, Pan, et al., 2018). Figure 2 further outlines three distinct regions with dotted lines that correspond to the ISC severity state in the plant. These states are defined based on the observed effects on the aforementioned health-related quantities: in the soft ISC region, these quantities exhibit minimal changes; in the moderate ISC region, these changes become more noticeable; in the severe ISC region, the changes are extremely significant.

3.2. Prognosis

To save memory space and computational costs, we only store the data points of the Gaussian posterior pdf estimate of the ISC state obtained from the EKF at the end of each cycle. This decision is made considering that the ISC state is not expected to change significantly within a single working cycle. Although the online ISC estimation algorithm is designed to acquire and store data every second, we prioritize saving only

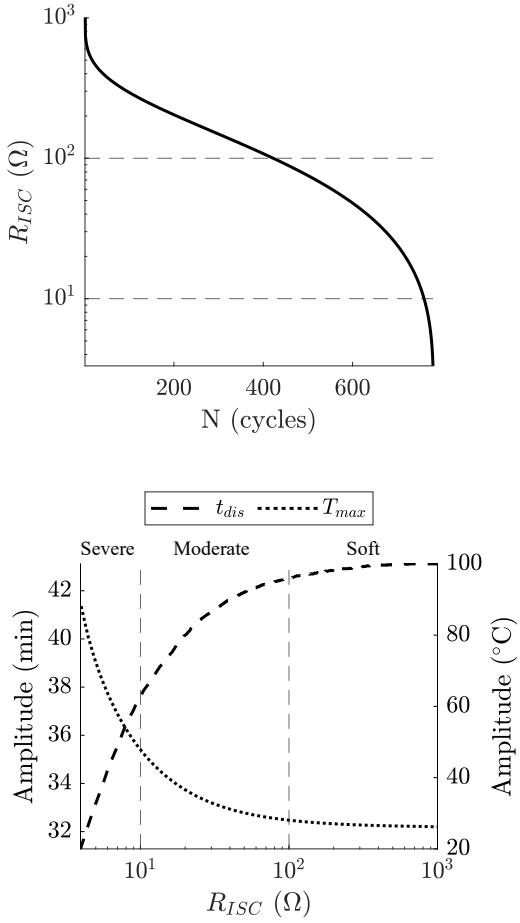


Figure 2. Top: ISC state evolution in the plant during cycling. Bottom: evolution of the maximum temperature T_{max} and the discharging time t_{dis} measured across one cycle.

these specific data points to local memory.

When enough data is available in the local memory, the user can request the prognosis. In this study, we define the prognosis triggering point as when the estimated ISC state enters the moderate ISC region, specifically when $\hat{R}_{ISC} \leq 100\Omega$, as indicated at the top of Figure 3. Additionally, we use a sliding window size W of 50 data points to ensure the linear regressor model captures the local trend behavior. This can be seen in the bottom part of Figure 3, which illustrates the linear regression at the prognosis query $N = 520$ cycles.

After fitting the parameters of the linear regressor model, the RUL pdf is computed. This is done by moving forward in time with the fitted model, sampling different realizations of the error terms from the estimated Gaussian distribution,

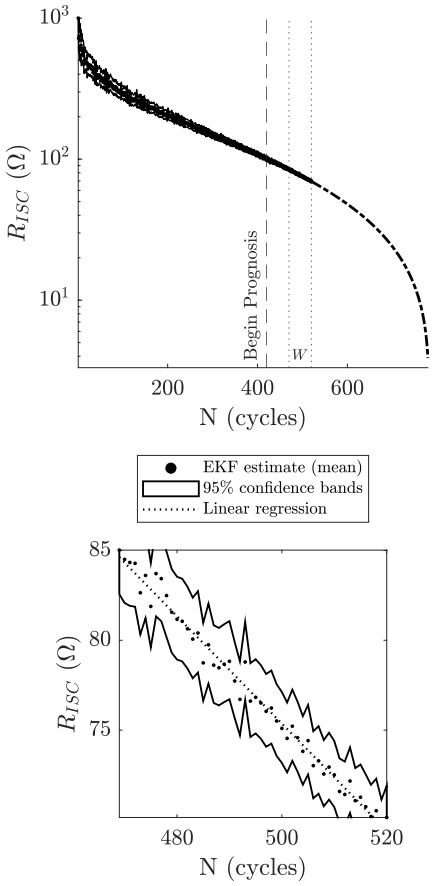


Figure 3. Top: Example of prognosis query when $N = 520$ cycles, outlining the sliding window W. Bottom: Linear regression on the latest available data points in the sliding window.

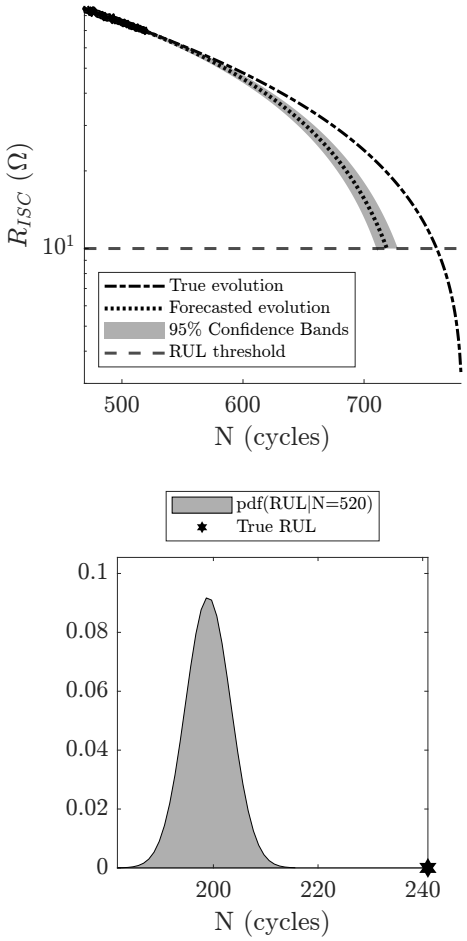


Figure 4. Top: predicted ISC state trajectory when $N = 520$ cycles. Bottom: corresponding RUL pdf estimation.

$\varepsilon_i \sim N(0, \sigma_\varepsilon)$, to account for modeling uncertainties. The prediction is then truncated up to a threshold value of 10 Ω to prevent the system from entering the severe ISC region. This process, at the prognosis query $N = 520$ cycles, is illustrated in Figure 4.

By repeating this process at different prognosis queries the RUL pdf prediction evolution is obtained, as shown in Figure 5, and compared with the actual RUL evolution of the plant. The results demonstrate the satisfactory performances obtained with the proposed method. The estimated RUL steadily converges to the true RUL value. However, the results consistently suggest that the true RUL is far outside the 95% confidence interval. This is because the linear regressor model can well approximate the local degradation behavior with good accuracy. However, the latter changes as ISC progresses due to the aforementioned exothermic electrochemical reactions, which are accounted for with the degradation model described in Eq. (13) considering an Arrhenius-like term. Furthermore, the confidence bounds narrow progressively as the prognosis steps advance, due to the increased accuracy of the EKF estimation as the ISC state becomes more severe, as also can be appreciated on top of Figure 3. This, in turn, leads to a smaller variance of the residuals with the estimated linear regressor at a later prognosis query, when the ISC state is more severe.

4. CONCLUSION

This work presents a prognosis framework for spontaneous ISC formation. The proposed framework leverages the potentialities of an EKF algorithm to online estimate and track the evolution of the equivalent ISC resistance value, which is a quantity representative of the actual ISC state of the battery cell. At appropriate instants, some of the estimated ISC resistance values are saved to local memory. The stored data are then processed for prognosis. When enough data are stored,

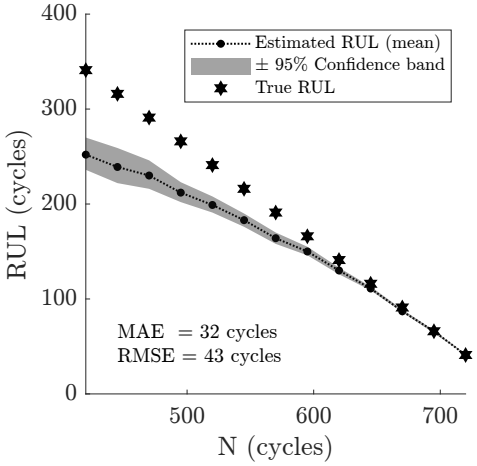


Figure 5. Comparison between true RUL evolution and predicted RUL evolution.

the user can request the prognosis algorithm to predict the battery RUL pdf. This is done by employing a linear regressor model for the prediction and a Monte Carlo simulation for quantifying the uncertainties involved. This work establishes a first step toward effective spontaneous ISC formation prognosis. Due to the lack of experimental data in the literature, the proposed approach has been validated numerically with synthetic measurements that aim to accurately reproduce the expected effects ISC has on the electrical and thermal characteristics of the cell. Furthermore, a degradation model that reasonably reproduces the expected evolution of ISC has been constructed based on certain assumptions that may not be fulfilled with real experimental data. Nevertheless, the framework proposed could in principle accommodate different degradation trajectories. Consequently, to further validate the approach, future studies will apply the methodology to real experimental data on spontaneous ISC formation as soon as the latter is available. On top of that, the method could be improved by using more sophisticated regressor models, such as ARIMA or NARX models, to improve the RUL prediction performance for a cell subjected to spontaneous ISC formation.

REFERENCES

Asakura, J., Nakashima, T., Nakatsuji, T., & Fujikawa, M. (2010, July 29). *Battery internal short-circuit detecting device and method, battery pack, and electronic device system.* Google Patents. (US Patent App. 12/670,597)

Cui, B., Wang, H., Li, R., Xiang, L., Zhao, H., Xiao, R., ... others (2024). Ultra-early prediction of lithium-ion battery performance using mechanism and data-driven fusion model. *Applied Energy*, *353*, 122080.

Ding, Y., Cano, Z. P., Yu, A., Lu, J., & Chen, Z. (2019). Automotive li-ion batteries: current status and future perspectives. *Electrochemical Energy Reviews*, *2*, 1–28.

Feng, X., Ouyang, M., Liu, X., Lu, L., Xia, Y., & He, X. (2018). Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. *Energy storage materials*, *10*, 246–267.

Feng, X., Pan, Y., He, X., Wang, L., & Ouyang, M. (2018). Detecting the internal short circuit in large-format lithium-ion battery using model-based fault-diagnosis algorithm. *Journal of Energy Storage*, *18*, 26–39.

Han, X., Lu, L., Zheng, Y., Feng, X., Li, Z., Li, J., & Ouyang, M. (2019). A review on the key issues of the lithium ion battery degradation among the whole life cycle. *ETransportation*, *1*, 100005.

Hermann, W. A., & Kohn, S. I. (2013, December 31). *Detection of over-current shorts in a battery pack using pattern recognition.* Google Patents. (US Patent 8,618,775)

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, *6*(9), 813-827.

Huang, L., Liu, L., Lu, L., Feng, X., Han, X., Li, W., ... others (2021). A review of the internal short circuit mechanism in lithium-ion batteries: Inducement, detection and prevention. *International Journal of Energy Research*, *45*(11), 15797–15831.

Ikeuchi, A., Majima, Y., Nakano, I., & KASA, K. (2014, July 3). *Circuit and method for determining internal short-circuit, battery pack, and portable device.* Google Patents. (US Patent App. 14/196,101)

Jia, Y., Brancato, L., Giglio, M., & Cadini, F. (2024). Temperature enhanced early detection of internal short circuits in lithium-ion batteries using an extended kalman filter. *Journal of Power Sources*, *591*, 233874.

Liu, H., Hao, S., Han, T., Zhou, F., & Li, G. (2023). Random forest-based online detection and location of internal short circuits in lithium battery energy storage systems with limited number of sensors. *IEEE Transactions on Instrumentation and Measurement*.

Ma, R., Deng, Y., & Wang, X. (2023). Simplified electrochemical model assisted detection of the early-stage internal short circuit through battery aging. *Journal of Energy Storage*, *66*, 107478.

Reichl, T., & Hrzina, P. (2018). Capacity detection of internal short circuit. *Journal of Energy Storage*, *15*, 345–349.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Simon, D. (2006). *Optimal state estimation: Kalman, h infinity, and nonlinear approaches*. John Wiley & Sons.

Yang, B., Cui, N., & Wang, M. (2019). Internal short circuit fault diagnosis for lithiumion battery based on voltage and temperature. In *2019 3rd conference on vehicle control and intelligence (cvci)* (pp. 1–6).

Yokotani, K. (2014, February 4). *Battery system and method for detecting internal short circuit in battery system.* Google Patents. (US Patent 8,643,332)

Zhang, G., Wei, X., Tang, X., Zhu, J., Chen, S., & Dai, H. (2021). Internal short circuit mechanisms, experimental approaches and detection methods of lithium-ion batteries for electric vehicles: A review. *Renewable and Sustainable Energy Reviews*, *141*, 110790.

Zheng, Y., Gao, W., Ouyang, M., Lu, L., Zhou, L., & Han, X. (2018). State-of-charge inconsistency estimation of lithium-ion battery pack using mean-difference model and extended kalman filter. *Journal of Power Sources*, *383*, 50–58.

BIOGRAPHIES

Lorenzo. BRANCATO was born in Italy on December 12, 1998. He earned his Bachelor's degree in Mechanical Engineering from Politecnico di Milano in 2020. Subsequently, he pursued a Master's degree in Mechatronic Engineering at Politecnico di Milano, completing his studies in 2022. He has started his Ph.D in Mechanical Engineering at Politecnico di Milano in 2023. His current research focus on the development of advanced diagnostic and prognostic approaches for dynamic, complex systems subject to degradation. This involves high-fidelity multi-physics modeling, simulation and model-based filtering methods.

Yiqi. JIA was born in China on January 28, 1996. She holds a Bachelor's degree in Automotive Engineering from Wuhan University of Technology, Wuhan, China (2017), and a Master's degree in Automotive Engineering from the University of Bath, Bath, UK (2018). After working as a vehicle engineer for nearly 3 years at Ford Motor Company, she started her Ph.D. journey in Mechanical Engineering at Politecnico di Milano in November 2021. Her research primarily focuses on the diagnosis and prognosis of Lithium-ion batteries, structural batteries, and more broadly, on mechanical/structural-related behaviors. This includes battery modeling, simulation, and data-based estimation methods for optimal battery management.

Marco. GIGLIO is Full Professor at the Department of Mechanical Engineering, Politecnico di Milano. His main research fields are: (i) Structural integrity evaluation of complex platforms through Structural Health Monitoring methodologies; (ii) Vulnerability assessment of ballistic impact damage on components and structures, in mechanical and aeronautical fields; (iii) Calibration of constitutive laws for metallic materials; (iv) Expected fatigue life and crack propagation behavior on aircraft structures and components; (v) Fatigue design with defects. He has been the coordinator of several European projects: HECTOR, Helicopter Fuselage Crack Monitoring and prognosis through on-board sensor, 2009-2011; ASTYANAX (Aircraft fuselage crack Monitoring System and Prognosis through eXpert onboard sensor networks), 2012-2015; SAMAS (SHM application to Remotely Piloted Aircraft Systems), 2018-2020. He has been the project leader of the Italian Ministry of Defence project in the National Plan of Military Research, SUMO (Development of a predictive model for the ballistic impact), 2011-2012, and SUMO 2 (Development of an analytical, numerical and experimental methodology for the design of ballistic multilayer protections), 2017-2019. He has published more than 210 papers, h-index 27 (source Scopus) in referred international journals and congresses.

Francesco. CADINI (MSc in Nuclear Engineering, Politecnico di Milano, 2000; MSc in Aerospace Engineering, UCLA, 2003; PhD in Nuclear Engineering, Politecnico di Milano, 2006) is Associate Professor at the Department of Mechanical Engineering, Politecnico di Milano. He has more than 20 years of experience in the assessment of the safety and integrity of complex engineering systems, entailing (i) artificial intelligence (machine learning)-based approaches for classification and regression, (ii) development and application of advanced Monte Carlo algorithms for reliability analysis (failure probability estimation), (iii) diagnosis and prognosis (HUMS) of dynamic, complex systems subject to degradation, (iv) uncertainty and sensitivity analyses, (v) structural reliability analyses.

Remaining Useful Lifetime Estimation of Bearings Operating under Time-Varying Conditions

Alireza Javanmardi[1,2], Osarenren Kennedy Aimiyekagbon[3], Amelie Bender[3], James Kuria Kimotho[4], Walter Sextro[3], and Eyke Hüllermeier[1,2]

[1] *Institute of Informatics, LMU Munich, Munich, Germany*

[2] *Munich Center for Machine Learning (MCML), Munich, Germany*

[3] *Chair of Dynamics and Mechatronics, Paderborn University, Paderborn, Germany*

[4] *Department of Mechanical Engineering, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya*

ABSTRACT

This paper investigates the remaining useful lifetime (RUL) estimation of bearings under dynamic, i.e., time-varying, operating conditions (OC). Unlike conventional studies that assume constant OC in bearing accelerated life tests, we introduce a dataset with time-varying OC during run-to-failure experiments, simulating real-world scenarios. We explore data-driven approaches to identify the transition point from a healthy to an unhealthy state and estimate the RUL. Additionally, we examine strategies for integrating OC information to enhance RUL estimations. These methodologies are evaluated through numerical experiments using various machine learning algorithms.

1. INTRODUCTION

Rolling element bearings are extensively used in industrial applications, such as wind turbines, electric motors, and generators. These bearings account for the largest percentage of failures in rotating machinery (Alewine & Chen, 2010) and about 40%-50% of all motor faults (Sharma et al., 2015). Failure of bearings may result in expensive downtime, increased maintenance costs due to failures propagating to other parts, and catastrophic effects if they support critical equipment. Predictive maintenance can be employed to increase the efficiency and reliability of bearings and technical systems in general, as it prevents unexpected failures and maximizes their availability. Predictive maintenance builds on prognostics, which involves the accurate estimation of the remaining useful lifetime (RUL) of technical systems or components, such as bearings.

For developing RUL estimation methods for bearings, existing datasets often focus on accelerated life tests conducted under constant operating conditions (OC) (Lee et al., 2007). In some cases where varying OC were taken into account, the OC only change between different run-to-failure experiments but remain constant within each experiment (Nectoux et al., 2012; Wang et al., 2018). However, this approach falls short of accurately simulating real-world scenarios where bearings may experience time-varying conditions throughout their operational life. Some research has been conducted on time-varying conditions. For example, Du et al. (2022) propose extracting features from the angular domain and RUL prediction based on the unscented particle filter. However, their proposed methodology was evaluated on ball bearing run-to-failure experiments, considering only varying rotating speed in the range [1450, 1550] rpm. Furthermore, the time point of degradation onset was manually determined. N. Li et al. (2019) propose a so-called "two-factor" state-space model based on a Wiener process, where the underlying degradation process is modeled in the state transition function, and the influence of the varying condition on the measured signal is captured in the measurement function of the proposed model. However, the proposed methodology builds on the assumption that the OC are known a priori and follow a known pattern. Furthermore, their methodology was evaluated on ball bearings subjected to cyclic varying speed conditions, taking up two speed values, namely 2200 rpm and 2600 rpm.

To address the presented limitation of existing studies and enhance the relevance to practical applications, we introduce a new dataset of bearing run-to-failure experiments, in which OC can dynamically vary over time, such as in a non-periodic and stochastic manner. Thus presenting new challenges for RUL estimation, as the vibration data not only reflects bearing degradation but is also influenced by changes in OC. Table 1 provides an overview of the existing datasets and their comparison to ours.

Table 1. Comparison between the publicly available datasets and our dataset.

	IMS (Lee et al., 2007)	Pronostia (Nectoux et al., 2012)	XJTU-SY (Wang et al., 2018)	LDM (Aimiyekagbon, 2024)
Bearing type	Roller bearing	Ball bearing	Ball bearing	Ball bearing
Rotating speed [rpm]	2000	1500, 1650, 1800	2100, 2250, 2400	$[1500, 3600]$
Static load [kN]	26.7	4, 4.2, 5	10, 11, 12	$[1.5, 4]$
Dynamic load amplitude [kN]	✗	✗	✗	$[0.5, 1.7]$
Dynamic load type	✗	✗	✗	Sinusoidal and Gaussian noise

Data-driven techniques for estimating the RUL involve establishing a mapping between available information, mainly vibration data, and the RUL. Typically, the initial step is to extract features from vibration data, given its high-dimensional nature and lack of a discernible trend for RUL estimation. Features are typically extracted in the time-, frequency-, and time-frequency-domain. While finding the best feature representation lies beyond the scope of this paper, we primarily adopt the Fast Fourier Transform (FFT) to obtain a frequency-domain representation of the vibration signal. Furthermore, to address the challenge of high dimensionality, we employ a methodology akin to that proposed in (Ren et al., 2018; von Hahn & Mechefske, 2022). This involves segmenting the resulting FFT signal into distinct frequency buckets and subsequently identifying the maximum value within each bucket.

It has been observed that the behavior of bearings does not exhibit a consistent trend from the beginning to the failure time. Instead, a typical scenario involves an initial phase of normal behavior followed by an abrupt shift at some point during the lifespan, indicating the initiation of degradation. These points, marking the transition from a healthy to an unhealthy state, are referred to as transition times. While existing approaches use various engineering techniques to detect these transition times (X. Li et al., 2019), this study employs a 2-means clustering technique on extracted features to define the transition time as the moment when the cluster of a bearing changes with respect to its initial cluster. After identifying the transition time, the subsequent data points can be used to train a RUL estimator model.

The dataset consisting of all features after the transition times, along with their corresponding RUL labels, can be fed into any supervised machine learning or deep learning model for fitting an RUL estimator. The challenge of estimating the RUL under dynamic OC is addressed through various approaches in the literature. Huang et al. (2019) incorporate the OC as an additional input in their deep network model. Fu et al. (2021) and Javanmardi & Hüllermeier (2023) suggest normalizing data according to OC. F. Li et al. (2020) integrate several algorithms into one model and select an optimal algorithm set for different OC to minimize their impact. Numerous studies address this problem by employing transfer learning or domain adaptation to handle the distribution shift between the training (source) and testing (target) domains (Mao et al., 2019; Fan et al., 2020; da Costa et al., 2020; Ding, Jia, Miao, & Huang, 2021; Ding, Jia, & Cao, 2021; Zhang et al., 2021). Ding et al. (2022) consider multi-source adaptation to manage the presence of subdomains in the source caused by multiple OC. To this end, we consider three distinct approaches in this study:

- Firstly, we train a regressor using only the previously attained features without taking the OC into account. This approach serves as a baseline for the subsequent two methods.
- Secondly, we employ the OC to normalize the features, aiming to mitigate its impact on the overall feature set.
- Thirdly, we concatenate the OC with the previously attained features, thereby incorporating them as additional features.

In the following sections, we first formalize the problem statement along with the details of all steps, from feature extraction to transition time determination and RUL estimation. Later, we elaborate on the data generation process and present comprehensive numerical results for the proposed approaches.

2. Problem Statement

Consider a dataset containing N instances of bearing run-to-failure data. Each bearing i in the dataset with a lifetime of T_i is represented as a time series $\mathbf{Z}_i := \left\{z_1^{(i)}, z_2^{(i)}, \ldots, z_{T_i}^{(i)}\right\}$. Here, $z_t^{(i)} := \left(v_t^{(i)}, o_t^{(i)}\right)$, where $o_t^{(i)} \in \mathbb{R}^{d_o}$ contains information about the operating and environmental conditions during the t^{th} measurement cycle, and $v_t^{(i)} \in \mathbb{R}^{d_v}$ represents the vibration signal collected during that measurement. For all $t \in [T_i] := \{1, \ldots, T_i\}$, the RUL $y_t^{(i)}$ of instance i at time t can be computed as follows:

$$y_t^{(i)} = T_i - t\,. \tag{1}$$

2.1. Feature Extraction from the Vibration Data

The vibration signal in the time domain $v_t^{(i)}$ is often high-dimensional, making it unsuitable for direct integration into a machine learning framework. In this study, we employ discrete Fourier transform to convert the signal into its frequency spectrum. This transformation results in $V_t^{(i)}$, a signal with the same dimensionality as the original time signal. Next, we partition the signal into m equally sized buckets $B_1, \ldots, B_m$ (with B_1 corresponding to the lowest frequency bucket and B_m to the highest) and simply extract the maximum ampli-

tude within each bucket to construct the m-dimensional frequency domain features $X_t^{(i)}$, i.e.,

$$X_t^{(i)} = \left(\max_{j_1 \in B_1} V_t^{(i)}(j_1), \ldots, \max_{j_m \in B_m} V_t^{(i)}(j_m) \right), \quad (2)$$

where $V_t^{(i)}(k)$ represents the k^{th} component of the signal $V_t^{(i)}$. Having access to this new feature, in the literature also known as the Spectrum-Principal-Energy-Vector (Ren et al., 2018), resolves the challenge posed by the high dimensionality of the initial vibration signal.

2.2. Transition Time Determination

K-means clustering is an unsupervised machine learning algorithm that clusters similar data points based on their proximity in the feature space. The algorithm initializes K cluster centroids and assigns each data point to the nearest centroid, recalculating the centroid of each cluster based on the mean of the assigned data points until convergence. The goal is to minimize the sum of squared distances between each data point and its assigned centroid. Here, we merely want to divide data points into a healthy or unhealthy cluster, thus K=2. We assume that each bearing starts in a healthy state, and hence, the cluster of the first point is considered healthy. A change in the cluster in the subsequent times is considered the beginning of the degradation. Once trained on the training data, the algorithm can be used in an online fashion for each test data instance to detect its changepoint promptly and initiate RUL prediction.

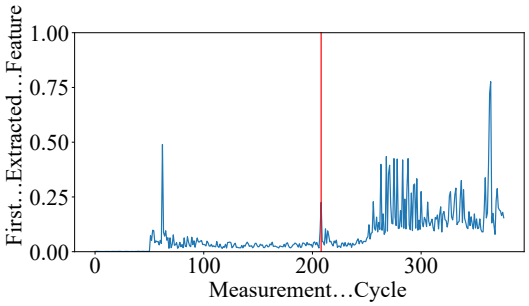
Following the extraction of low-dimensional features $X_t^{(i)}$ from the vibration data, we can utilize a 2-means clustering algorithm to assign a cluster label $\delta_t^{(i)} \in \{0, 1\}$ to each measurement time t for every bearing i. Subsequently, we define $t_{\text{TT}}^{(i)}$, the *transition time*, as the moment when the cluster of the i^{th} bearing differs from its initial cluster. Formally, this is expressed as

$$t_{\text{TT}}^{(i)} = \min \left\{ t : t \in [1 + T_c, T_i] \text{ and } \delta_t^{(i)} \neq \delta_1^{(i)} \right\}, \quad (3)$$

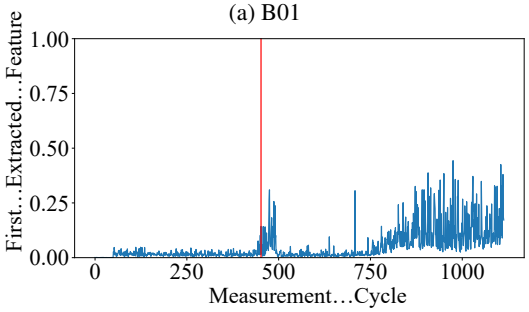
where T_c serves as a hyperparameter, representing the tolerance level. It signifies that a change in the cluster occurring earlier than T_c is not considered in the transition time calculation. The transition times for two bearing experiments are exemplarily depicted in Figure 1. Once the transition times are determined, we can define healthy and unhealthy datasets as follows:

$$\mathcal{D}_{\text{healthy}} = \left\{ (X_t^{(i)}, o_t^{(i)}, y_t^{(i)}) : i \in [N], t < t_{\text{TT}}^{(i)} \right\}, \quad (4)$$

$$\mathcal{D}_{\text{unhealthy}} = \left\{ (X_t^{(i)}, o_t^{(i)}, y_t^{(i)}) : i \in [N], t \geq t_{\text{TT}}^{(i)} \right\}. \quad (5)$$



(a) B01



(b) B04

Figure 1. The first extracted feature (blue) plotted for two bearings alongside their determined transition times (red). Here T_c is set to 150.

2.3. RUL Estimation

After extracting features from the vibration data and determining transition times, the next step is to estimate the RUL of the bearing. The primary focus of this paper is to leverage machine learning algorithms for that purpose. From a machine learning perspective, the problem is framed as a supervised regression setting—finding a mapping from the feature space to the RUL space. However, we have yet to explore how to benefit from OC information. In this context, we consider three distinct scenarios as outlined below and depicted in the flowchart in Figure 2.

- **Scenario 1 (disregarding OC):** In this scenario, OC information is neglected, and training proceeds without considering such contextual data.
- **Scenario 2 (OC for feature scaling):** This approach involves utilizing OC information for data/feature normalization. The methodology employs PCA to reduce the dimensionality of OC data from d_o to 1. Subsequently, a uniform discretization method is applied to bin the resulting one-dimensional feature into B bins. Next, inspired by a prior study (Javanmardi & Hüllermeier, 2023), the data in each bin is normalized to the $[0, 1]$ interval using B distinct MinMax scalers, aiming to mitigate the impact of diverse OC indirectly.
- **Scenario 3 (OC as additional features):** In this method, OC information is treated as an additional set of features, thereby augmenting the feature space. The objective is to

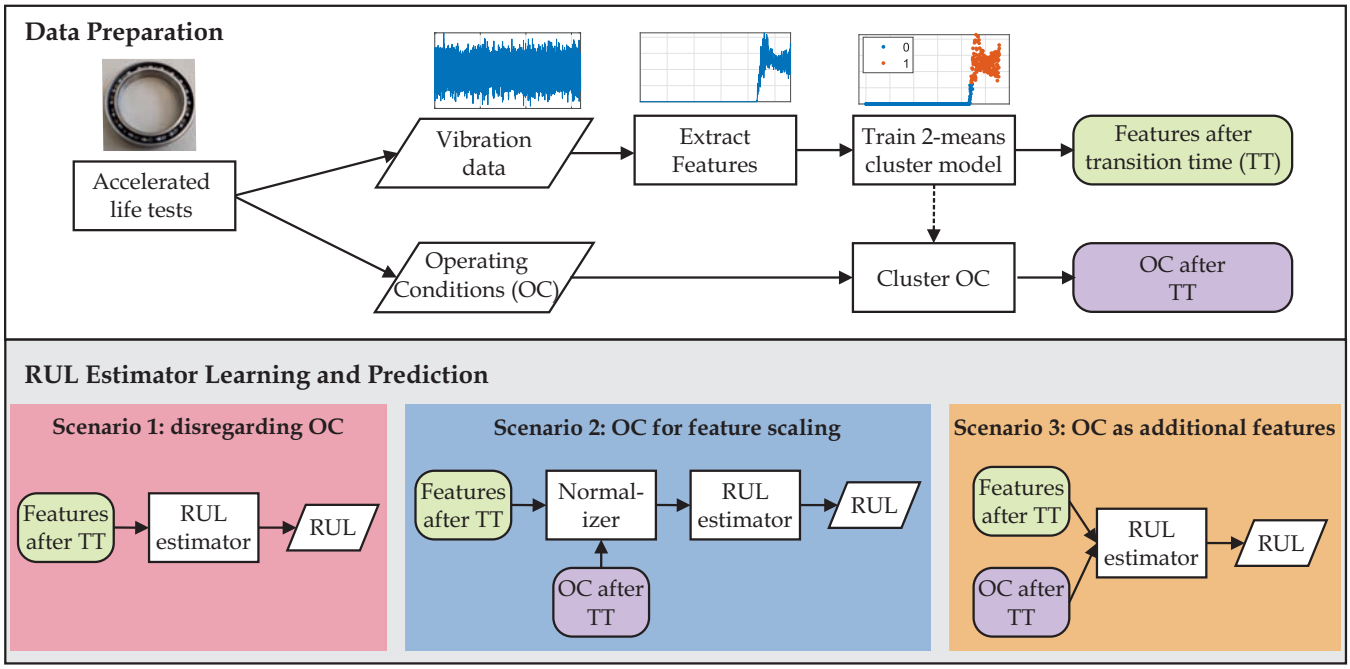


Figure 2. Flow chart of the proposed method.

enable the machine learning model to identify and consider interactions between OC and vibration features during the RUL estimation process.

Note that any machine learning or deep learning model can be used as the underlying RUL estimator for the three proposed scenarios. In this paper, we focus on traditional machine learning models, such as gradient boosting (GB) and random forest (RF).

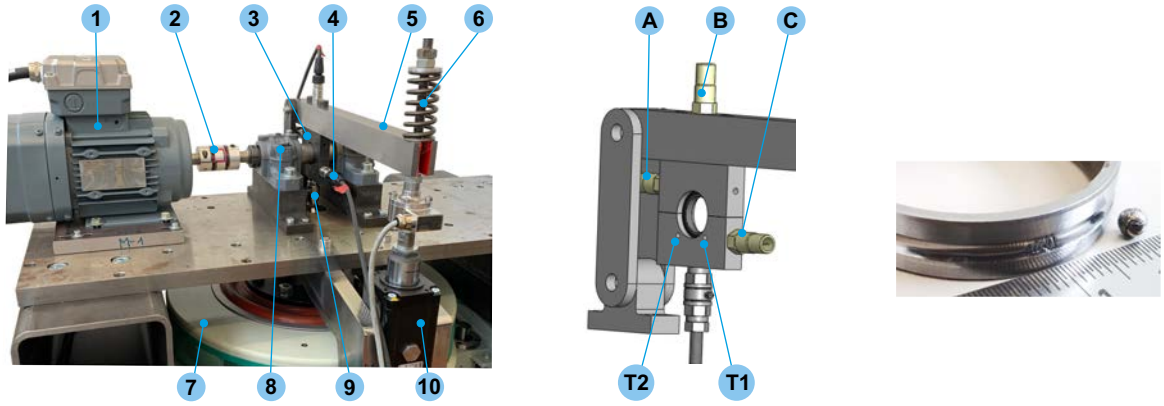
3. Case Study

The experimental dataset, which consists of accelerated life tests of ball bearings subjected to time-varying conditions, is gathered at the Chair of Dynamics and Mechatronics (LDM) at Paderborn University. The specifications of the test bearing allow an experiment with valuable condition monitoring data to take several hours. Specifically, the 61806-2RS rolling element bearing with a basic static load rating $C_0 = 3.15\,\mathrm{kN}$ and a dynamic load rating $C = 4.00\,\mathrm{kN}$ have a basic rating life L_{10} of approximately five hours while considering a constant equivalent load of 4.50 kN, a rotating speed of 2500 rpm and other factors not been considered, such as lubrication.

The bearing test rig with its components is captured in Figure 3(a). The test bearing within its housing (3) is mounted on a shaft. The shaft is coupled with the driving motor (1) via a jaw coupling (2) and supported by two spherical roller bearings (8) within their housing. A static pre-load is exerted on the bearing via a lever structure (5), which is attached to the bearing housing. To this end, the compression spring, mounted on the lever structure, is compressed by the linear actuator (10). A dynamic load is superimposed on the static pre-load by means of an electrodynamic shaker (7), which is connected to the test bearing housing via a stinger (9).

The input signals, namely the exerted forces and shaft rotating speed, are measured synchronously with vibration and temperature as condition monitoring data. Three one-directional accelerometers (4) measure the vibration of the bearing indirectly. Two accelerometers (A and C) measure the vibration horizontally from the housing, and one (B) measures vertically from the lever structure, as illustrated in Figure 3(b). The ambient temperature and bearing temperature are measured with Pt100 resistance thermometers. The bearing temperature is measured indirectly from its housing at the positions (T1 and T2) depicted in Figure 3(b). Measurements were acquired at a sampling duration of 1.6 s and a measurement interval of approximately 12 s. The temperature signals are measured with a sampling rate of 10 Hz. To facilitate high-frequency analysis, vibration data were sampled at 128 kHz for experiments till B09 and due to data storage issues at 64 kHz for experiments from B10. This lower sampling frequency is theoretically sufficient for analysis in the frequency range of interest up to 32 kHz.

During an experiment, the test bearing is subjected to dynamic load superimposed on a static pre-load. To accommodate different dynamic load types, the dynamic load is sinusoidal with a constant frequency of 2 Hz for some experiments. The amplitude of the sinusoidal load is stationary per measurement and takes on a random value from a station-



(a) (b) (c)

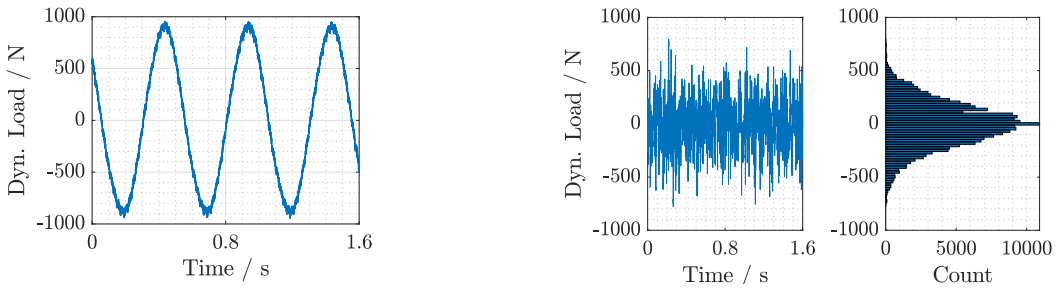
Figure 3: (a) Ball bearing test-rig with the following components: (1) motor, (2) jaw coupling, (3) bearing housing, (4) accelerometers, (5) lever structure, (6) electrodynamic shaker (DFG, 2017), (7) support bearing housing, (8) stinger connected to a quartz force sensor, and (9) linear actuator.
(b) Accelerometer and temperature placement on the test bearing housing without the shaft
(c) A dismantled test bearing with a surface defect on the inner ring raceway and spalls on a rolling element.

ary uniform distribution within a predefined interval between measurements. For other experiments, the dynamic load is Gaussian white noise with a maximum excitation frequency of 200 Hz and truncated to remain within a predefined interval between measurements. A measurement of the dynamic load types is exemplarily shown in Figure 4. The shaft rotating speed is also set to be constant per measurement and takes on a random value from a stationary uniform distribution within a predefined interval between measurements. The predefined range of values per experiment can be found in Table 2.

To avoid failure of other components, except the test bearing, an experiment ends when the test bearing fails. Bearing failure is determined by two predetermined failure threshold criteria. If one threshold is exceeded, the experiment is stopped. On the assumption that the test bearing is in a normal state, without previous loading history, and the ranges of speed and force are restrained, the first criterion builds on the equivalent energy content of the vibration data and is formulated as:

$$F_{vibration} = O \cdot \frac{1}{m} \sum_{t=1}^{m} \text{RMS}\left(v_t^{(i)}(1), \ldots, v_t^{(i)}(n)\right), \quad (6)$$

where $F_{vibration}$ denotes the vibration threshold value, $O = 8$ is a predetermined constant value, m is the number of vibration signals to consider, and $v_t^{(i)}(k)$ represents the k^{th} index of a vibration signal $v_t^{(i)}$ of length n. For the provided experiments, this threshold value lies approximately between 6 g and 10 g. Since improper lubrication of the bearing raceway leads to increased friction and subsequently to increased temperature, the second criterion builds on the bearing temperature. According to the data sheet and to avoid melting the bearing seal made up of nitrile butadiene rubber (NBR), the threshold value, based on the bearing housing temperature $F_{temperature}$, is set as 110 °C.



(a) (b)

Figure 4. Dynamic load: (a) Sinusoidal load and (b) Gaussian white noise.

Table 2. Set range of OC for experiments.

Experiment	Rotating speed [rpm]	Static load [N]	Dynamic load [N]	Dynamic load type
B01 - B03	$[2400, 3000]$	$[3300, 3800]$	$[500, 1400]$	Sinusoidal
B04 - B05	$[1500, 3000]$	$[2500, 3800]$	$[500, 1500]$	Sinusoidal
B06	$[1500, 3600]$	$[3300, 3800]$	$[750, 1700]$	Sinusoidal
B07	$[1500, 3000]$	$[3250, 4000]$	$[250, 750]$	Gaussian noise
B08	$[1500, 3000]$	$[3250, 4000]$	$[500, 1000]$	Gaussian noise
B09	$[1500, 3000]$	$[2500, 3800]$	$[750, 1000]$	Gaussian noise
B10	$[1500, 2700]$	$[2000, 3250]$	$[750, 1500]$	Gaussian noise
B11 - B13	$[1500, 2700]$	3000	1000	Gaussian noise
B14 - B15	2700	2500	$[750, 1500]$	Gaussian noise
B16	2700	2500	$[1000, 1500]$	Gaussian noise
B17	2700	$[1500, 2500]$	1500	Gaussian noise

In this paper, 17 run-to-failure experiments are utilized, with the temperature threshold value $F_{temperature}$ being exceeded for four experiments (B03, B07, B08, and B15) and the vibration threshold value $F_{vibration}$ for others. The failure types are not predetermined, but several single or combined failure types, such as an outer ring raceway defect or rolling element fault combined with an inner ring raceway defect, could occur during an experiment. Figure 3(c) is an image of a dismantled test bearing with spalls on a rolling element and an inner ring raceway defect after an experiment (B06). Due to the proximity to the bearing and for brevity, only the horizontal accelerometer (labeled A) is exemplarily considered in the following analysis. Also, three-dimensional OC information is considered, including peak dynamic load [N], mean absolute static load [N], and mean absolute rotating speed [rpm].

3.1. Numerical Experiments

We set the bucket size for feature extraction at 20. To conduct fair experiments, we repeat the proposed methods 17 times, reserving one bearing for testing each time while using the data of the remaining bearings for training. Without loss of generality, take the j^{th} bearing as the test bearing. The training and test data are defined as follows:

$$\mathcal{D}^{\text{train}} = \left\{ (X_t^{(i)}, o_t^{(i)}, y_t^{(i)}) : i \in [N] \setminus \{j\}, t \in [T_i] \right\}, \quad (7)$$

$$\mathcal{D}^{\text{test}} = \left\{ (X_t^{(j)}, o_t^{(j)}, y_t^{(j)}) : t \in [T_j] \right\}. \quad (8)$$

We then define X_{train} as the collection of features from all bearings in the training data. This data is normalized and fed into the 2-means clustering algorithm. We found out that utilizing only the first ten features is sufficient for clustering, yielding stable transition times. For the sake of fair comparison, the same transition times are used in all three RUL estimation scenarios. The resulting transition points lead to the creation of a new training dataset, which consists of only the data after the transition times, aka *unhealthy* points:

$$\mathcal{D}^{\text{train}}_{\text{unhealthy}} = \left\{ \left(X_t^{(i)}, o_t^{(i)}, \tilde{y}_t^{(i)} \right) : i \in [N] \setminus \{j\}, t \geq t_{\text{TT}}^{(i)} \right\}, \quad (9)$$

where

$$\tilde{y}_t^{(i)} = \frac{T_i - t}{T_i - t_{\text{TT}}^{(i)}} \times 100\% \quad (10)$$

is the RUL percentage after the transition time.

Different RUL estimation scenarios require different input data, as illustrated in Figure 2. The details are provided as follows.

- Scenario 1: For this approach, we simply use the training data in the form

$$\left\{ \left(X_t^{(i)}, \tilde{y}_t^{(i)} \right) : i \in [N] \setminus \{j\}, t \geq t_{\text{TT}}^{(i)} \right\} . \quad (11)$$

- Scenario 2: Here, we utilize the training OC for PCA and divide its one-dimensional output space into 20 bins. This way, each $o_t^{(i)} \in \mathbb{R}^3$ is replaced with its discretized counterpart $\tilde{o}_t^{(i)} \in [20]$. Next, for each region $r \in [20]$, a distinct normalizer MinMax_r is applied to the features with the same operating region, i.e.,

$$X_{\text{train}}^r := \left\{ X_t^{(i)} : i \in [N] \setminus \{j\}, t \geq t_{\text{TT}}^{(i)}, \tilde{o}_t^{(i)} = r \right\} . \quad (12)$$

  Let $\tilde{X}_t^{(i)} := \text{MinMax}_{\tilde{o}_t^{(i)}}(X_t^{(i)})$ be the normalized counterpart of $X_t^{(i)}$. The training data for this method can be written as

$$\left\{ (\tilde{X}_t^{(i)}, \tilde{y}_t^{(i)}) : i \in [N] \setminus \{j\}, t \geq t_{\text{TT}}^{(i)} \right\} . \quad (13)$$

- Scenario 3: We simply concatenate feature vectors and OC to create 23-dimensional features. The training data would be in the form

$$\left\{ \left([X_t^{(i)}; o_t^{(i)}], \ \tilde{y}_t^{(i)} \right) : i \in [N] \setminus \{j\}, t \geq t_{\text{TT}}^{(i)} \right\} . \quad (14)$$

We made the data, as well as all the implementations, publicly available on Zenodo[1] and GitHub[2] to encourage further development of RUL estimation models in dynamic operating conditions.

3.2. Results and Discussion

We employed GB and RF as models for estimating RUL. For each test bearing, separate models were trained for every estimation scenario. Figure 5 compares the performance of the three RUL estimation scenarios for two different bearings. To mitigate random effects, model training was repeated ten times for each test bearing and each scenario, using ten different random seeds for both GB and RF, and the resulting average mean absolute error (MAE) is reported in Table 3. Notably, the standard deviation of the MAE values was negligible and, therefore, not included in the table. The best MAE value for each bearing scenario is highlighted in bold.

The findings showcased in Table 3 shed light on the potential benefits of integrating OC information in the context of RUL estimation. Both learning models exhibited a decrease in MAE for more than 50% of the bearings when OC details were taken into account (scenarios 2 and 3 combined), with

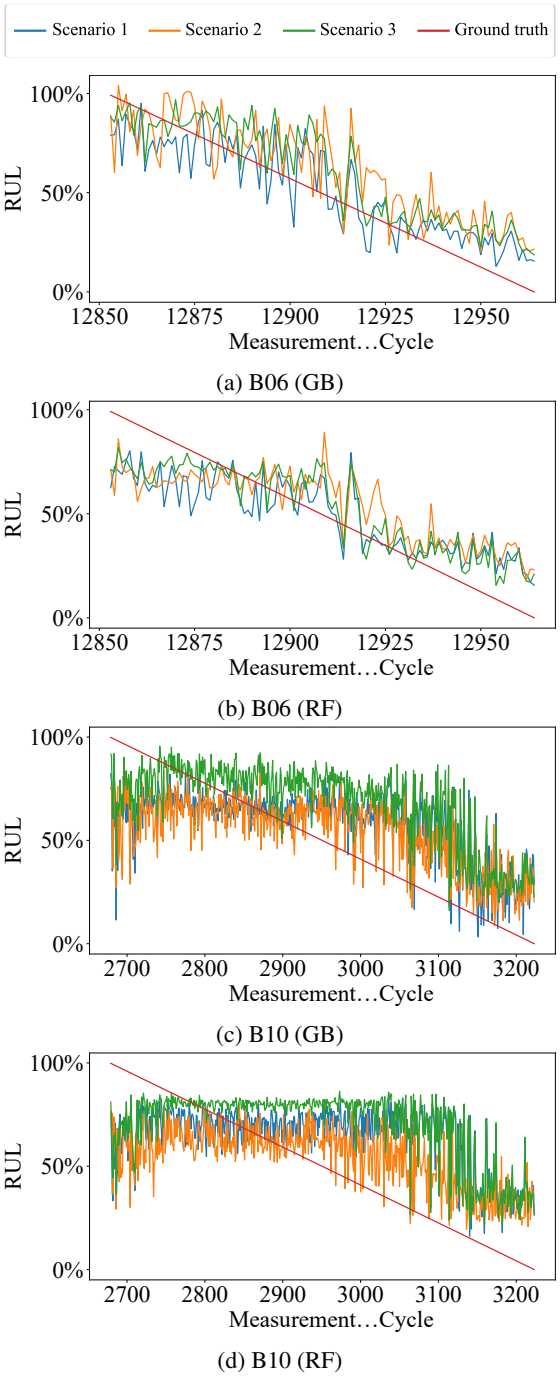


Figure 5. RUL estimation performance comparison for three scenarios.

RF benefiting more compared to GB. It should be noted that preprocessing steps, such as feature extraction and transition time identification, can also affect the final outcomes. Despite this, the primary focus here is to compare the performance of different scenarios under fixed preprocessing steps.

[1] https://doi.org/10.5281/zenodo.10805042
[2] https://github.com/alireza-javanmardi/bearing-RUL

Table 3. MAE of the predictions in different scenarios.

			GB			**RF**		
Bearing	Total lifetime	Transition time	Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
B01	377	208	**18.14**	22.77	19.96	18.00	21.51	**17.17**
B02	1116	998	29.43	**25.25**	27.41	28.74	**26.10**	27.24
B03	614	562	24.05	**21.52**	25.83	25.24	**21.90**	24.17
B04	1114	452	15.23	19.25	**15.02**	14.90	19.56	**14.07**
B05	572	560	44.40	41.11	**36.34**	42.20	37.28	**34.75**
B06	12965	12853	**11.73**	16.00	13.55	13.50	15.64	**11.85**
B07	6393	6205	44.36	**42.54**	42.82	43.63	44.78	**43.09**
B08	1827	1219	**15.42**	17.98	15.70	18.18	19.08	**17.51**
B09	1813	253	20.85	23.19	**17.19**	21.41	22.95	**19.67**
B10	3224	2679	19.48	**18.62**	23.13	23.87	**18.79**	26.11
B11	1953	931	23.94	23.85	**22.79**	25.22	**24.28**	24.66
B12	767	154	**15.80**	17.26	17.43	16.90	**16.27**	16.93
B13	19417	18022	26.91	**25.69**	27.36	**24.77**	26.78	29.13
B14	12317	12050	30.71	**26.75**	30.10	30.14	**29.10**	30.36
B15	22567	21051	16.74	19.88	**14.17**	14.20	21.32	**13.66**
B16	5891	5400	**20.17**	21.42	24.74	20.28	**18.41**	21.74
B17	2733	2323	22.19	**21.88**	24.47	22.37	24.15	**21.58**

4. Summary and Outlook

To address the limitation of existing studies and enhance the relevance to practical applications, a new ball bearing run-to-failure dataset, considering time-varying operating conditions (OC), is introduced. Specifically, during an experiment, the test bearing is subjected to a sinusoidal load or Gaussian white noise superimposed on a static pre-load. Furthermore, the rotating speed takes on a random value from a stationary uniform distribution within a predefined interval between measurements. Owing to the degradation path of the ball bearings, a 2-means clustering algorithm is employed to partition the features extracted from raw vibration data into two states, namely healthy and unhealthy states. To estimate the remaining useful lifetime (RUL) for the unhealthy state, even under such time-varying OC, three different scenarios are considered, namely, **Scenario 1**, where the measured OC are disregarded, **Scenario 2**, where the OC are employed for feature scaling, and **Scenario 3**, where the OC serve as auxiliary features. Different machine learning techniques, such as gradient boosting and random forest, are employed as the RUL estimator for each scenario. The results of the presented case study suggest that the usefulness of incorporating OC information depends on the individual case: in some scenarios, it is clearly advantageous, and in others, it does not yield significant benefits.

As a future work, one may delve deeper into other ways of incorporating OC information into RUL estimation. For instance, a hybrid model consisting of a physics-based model and a machine learning model can be an interesting extension. The physics-based model could capture the relationship between the varying OC and the system state, while the machine learning model could capture the relationship between the measured system parameters and the RUL. Moreover, more advanced learning algorithms, such as deep learning techniques, along with tools from domain adaptation and transfer learning, can be employed on this dataset to determine whether they can enhance the results.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Project number 451737409.

References

Aimiyekagbon, O. K. (2024). *Run-to-failure data set of ball bearings subjected to time-varying load and speed conditions.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.10805042` (Data set)

Alewine, K., & Chen, W. (2010). Wind turbine generator failure modes analysis and occurrence. In *Wind power 2010 conference, dallas.*

da Costa, P. R. d. O., Akçay, A., Zhang, Y., & Kaymak, U. (2020). Remaining useful lifetime prediction via deep domain adaptation. *Reliability Engineering & System Safety*, *195*.

DFG. (2017). *Schwingungsanalysesystem für automatisierte Schwingungsmessungen. Major research instrumentation supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Project number 391178551*.

Ding, Y., Ding, P., Zhao, X., Cao, Y., & Jia, M. (2022). Transfer learning for remaining useful life prediction across operating conditions based on multisource domain adaptation. *IEEE/ASME Transactions on Mechatronics*, *27*.

Ding, Y., Jia, M., & Cao, Y. (2021). Remaining useful life estimation under multiple operating conditions via deep subdomain adaptation. *IEEE Transactions on Instrumentation and Measurement*, *70*.

Ding, Y., Jia, M., Miao, Q., & Huang, P. (2021). Remaining useful life estimation using deep metric transfer learning for kernel regression. *Reliability Engineering & System Safety*, *212*.

Du, W., Hou, X., & Wang, H. (2022). Time-varying degradation model for remaining useful life prediction of rolling bearings under variable rotational speed. *Applied Sciences*, *12*. Retrieved from https://www.mdpi.com/2076-3417/12/8/4044

Fan, Y., Nowaczyk, S., & Rögnvaldsson, T. (2020). Transfer learning for remaining useful life prediction based on consensus self-organizing models. *Reliability Engineering & System Safety*, *203*.

Fu, S., Zhong, S., Lin, L., & Zhao, M. (2021). A novel time-series memory auto-encoder with sequentially updated reconstructions for remaining useful life prediction. *IEEE Transactions on Neural Networks and Learning Systems*, *33*.

Huang, C.-G., Huang, H.-Z., & Li, Y.-F. (2019). A bidirectional lstm prognostics method under multiple operational conditions. *IEEE Transactions on Industrial Electronics*, *66*.

Javanmardi, A., & Hüllermeier, E. (2023). Conformal prediction intervals for remaining useful lifetime estimation. *International Journal of Prognostics and Health Management*, *2*.

Lee, J., Qiu, H., Yu, G., Lin, J., & Services, R. T. (2007). IMS, university of cincinnati. bearing data set. *NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA*.

Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., … Huang, Z. (2020). An optimal stacking ensemble for remaining useful life estimation of systems under multi-operating conditions. *IEEE Access*, *8*.

Li, N., Gebraeel, N., Lei, Y., Bian, L., & Si, X. (2019). Remaining useful life prediction of machinery under time-varying operating conditions based on a two-factor state-space model. *Reliability Engineering & System Safety*, *186*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0951832018313024

Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability engineering & system safety*, *182*, 208–218.

Mao, W., He, J., & Zuo, M. J. (2019). Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, *69*.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. In *Ieee international conference on prognostics and health management, phm'12.* (pp. 1–8).

Ren, L., Sun, Y., Wang, H., & Zhang, L. (2018). Prediction of bearing remaining useful life with deep convolution neural network. *IEEE access*, *6*, 13041–13049.

Sharma, S., Abed, W., Sutton, R., & Subudhi, B. (2015). Corrosion fault diagnosis of rolling element bearing under constant and variable load and speed conditions. *IFAC-PapersOnLine*, *48*.

von Hahn, T., & Mechefske, C. K. (2022). Knowledge informed machine learning using a weibull-based loss function. *Journal of Prognostics and Health Management*, *2*.

Wang, B., Lei, Y., Li, N., & Li, N. (2018). A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, *69*(1), 401–412.

Zhang, W., Li, X., Ma, H., Luo, Z., & Li, X. (2021). Transfer learning using deep representation regularization in remaining useful life prediction across operating conditions. *Reliability Engineering & System Safety*, *211*.

Residual Selection for Observer-Based Fault Detection and Isolation in a Multi-Engine Propulsion Cluster

Renato Murata[1,2], Julien Marzat[1], Hélène Piet-Lahanier[1], Sandra Boujnah[2], and Pierre Belleoud [2]

[1] *DTIS ONERA, Université Paris-Saclay, Palaiseau, 91120, France*
firstName.lastName@onera.fr

[2] *CNES, Sous-Direction Techniques Systèmes de Transport Spatial, Paris, 75012, France*
firstName.lastName@cnes.fr

ABSTRACT

For complex systems, the number of residual candidates generated by Structural Analysis could be in the order of tens of thousands, and implementing all candidates is infeasible. This paper addresses the residual generator candidate selection problem from a state-observer perspective. First, the most suitable candidates to derive state-observers are selected based on two criteria related to the state-space form and a low number of equations. Then, a novel algorithm finds the minimal subset of residual generator candidates capable of detecting and isolating all faults. A procedure is introduced to compare the fault sensitivity of the selected candidates. This residual selection method is applied to the multi-engine propulsion cluster of a reusable launcher to illustrate its benefits.

1. INTRODUCTION

A classical model-based approach for fault detection and isolation usually comprises two main steps: the residual generation and the residual evaluation (Simani et al., 2003). The first step relies on the mathematical model of the system to generate signals, called residuals, that contain fault information. Then, the presence of the faults is inferred by a residual evaluation method. Structural Analysis (SA) has been proven to be a powerful tool for developing model-based fault diagnosis systems (Escobet, Bregon, Pulido, & Puig, 2019). It is a graph-based tool that uses the model equations to build a structural model. From the structural model, efficient algorithms (Krysander, Åslund, & Nyberg, 2007) can be applied to find residual generator candidates automatically. However, the number of candidates increases exponentially with the number of sensors. For large-scale systems, the number of residual generator candidates can be in the order of tens of thousands. This brings a new problem to be solved: the selection of the best subset of residuals that meets both fault detectability and isolability requirements.

The residual selection problem is addressed in (Svärd, Nyberg, & Frisk, 2013) where algorithms are proposed to find a minimal subset of residuals to meet the isolability constraints. In (Jung & Frisk, 2018), the residual selection problem is solved using convex optimization. In this case, the optimization problem depends on recorded data to find the minimal and most effective subset of residuals. In (Jung & Sundström, 2017), the residual selection problem is addressed by combining the fault sensitivity information of the residuals with machine learning methods. However, in all of those works, the fault isolability constraint used is very restrictive, leading to sub-optimal solutions with more residuals than necessary to isolate all faults.

Here, it is proposed to use a different fault isolability constraint based on the fault signature. This less restrictive constraint is able to lead to an optimal subset of residuals with minimal cardinality. Such an isolability constraint based on the fault signature has been used previously in (Zhang & Rizzoni, 2017) for residual selection. However, the objective was to find a subset of residuals that would produce the most "unique" fault signature for robustness purposes.

This paper proposes a new algorithm to find the minimal subset of residual generators able to detect and isolate predefined faults. The algorithm is adapted for an observer-based residual generation technique. State observers are more robust to modeling errors and parameter uncertainty when compared with other model-based residual generation techniques, such as Sequential Residual Generation (Isermann, 2005). The idea is to select the residual generator candidates based on

two main criteria: the candidates that can be easily written into the state-space form and the residual generators with the lowest number of equations. The state-space form is required to implement the majority of observers, such as the Kalman filter (Kalman, 1960) or Luenberger observers (Luenberger, 1964). It is preferred to have fewer equations because each one has a degree of uncertainty and modeling errors.

Depending on the number of residual generator candidates, it is possible to find many subsets of residuals with minimal cardinality. In order to choose the most suitable residuals in terms of fault sensitivity, a procedure based on the equations of the residual generator candidates is proposed. It quantifies the impact that a fault will have on the measured variables.

The main contributions of this paper are as follows. First, an algorithm to find the minimal subset of residuals to detect and isolate all faults. Second, a procedure based on the equations of the residual generators is proposed to compare the sensitivity of the residuals for one specific fault. The paper is organized as follows. In Section 2, basic notions of model-based diagnosis are recalled. In Section 3, the minimal residual selection problem is described and an algorithm to solve this problem is proposed in Section 4. Section 5 describes the procedure that uses the residual generator equations to compare the sensitivity of two residual generator candidates for a given fault. In Section 6, the proposed algorithm is applied in a multi-engine propulsion cluster of a reusable launcher. Conclusions are presented in Section 7.

2. Preliminaries on Model-Based Diagnosis

This section recalls some model-based diagnosis notions needed to formulate the residual selection problem formally introduced in (Svärd et al., 2013). Those notions are used to define necessary conditions to meet detectability and isolability constraints. Consider a model defined as

$$M = (E, X, Z, F) \tag{1}$$

where E is the vector of n_e system equations, X the vector of unknown variables in $\mathbb{R}^{n_x}$, Z the vector of known variables in $\mathbb{R}^{n_z}$ and F the vector of fault variables in $\mathbb{R}^{n_f}$. It is assumed that each fault $f \in F$ affects only one equation $e \in E$. This basic assumption is not as limiting as it may initially appear, as the equation e affected by the fault can propagate its effect through other equations. If a fault affects simultaneously more than one equation in the system, the system may be poorly modeled. Given the model (1), an ideal residual generator is defined as

Definition 2.1 (Ideal residual generator) *Consider a model M such as* (1)*. A system R with input Z and output r is a residual generator for M, and r is a residual if $f = 0$ implies $r = 0$ for all $f \in F$.*

In reality, residuals slightly deviate from zero even when no fault is present in the system due to unmodeled dynamics such as measurement noise and parameter uncertainty. One important property of residuals is their fault sensitivity, which defines the subset of faults that will affect this residual:

Definition 2.2 (Fault sensitivity) *Let R_i be a residual generator for model M. Then R_i is sensitive to fault $f \in F$ if $f \neq 0$ implies $r_i \neq 0$.*

With a set of residual generators $R \supseteq R_i, i \in \mathbb{N}$, the fault signature S_f of a fault f can be defined. The fault signature describes the subset of residuals that are sensitive to this fault:

Definition 2.3 (Fault signature) *For a set of residual generators R, the fault signature S_f of a fault f contains all the residuals $R_f \subseteq R$ sensitive to f.*

Using the fault signature, the fault isolability can be defined. If the fault has a unique signature, i.e., a unique subset of residuals is sensitive to it, the fault can be isolated from the others.

Definition 2.4 (Fault signature isolability) *A fault f is isolable using a set of residual generators R if its fault signature S_f is unique when compared to the other fault signatures.*

3. Minimal Residual Selection Problem

The minimal residual selection problem is formally defined as an optimization problem. Considering all residual generators available R_{all} to detect and isolate n_f faults, the objective is to find a minimal subset of R_{all} that respects the fault signature isolability property presented in def. 2.4, i.e., that generates unique fault signatures S_i for each fault $f_i, i = 1, 2..., n_f$. The optimization problem is formulated as

$$\begin{aligned} \min_{R \subseteq R_{all}} \quad & |R| \\ \text{s.t.} \quad & S = \{S_1, S_2, ..., S_{n_f}\}_{\neq} \\ & S \neq 0 \end{aligned} \tag{2}$$

where $|R|$ is the cardinality of the subset R. An equivalent optimization problem, using the fault signature, is introduced in (Zhang & Rizzoni, 2017), but a solution to this problem is not addressed.

The fault signature isolability concept is a key notion of finding the minimal subset of residuals to isolate all the faults. In previous works, such as (Svärd et al., 2013) and (Jung & Frisk, 2018), a different fault isolability definition was used. For instance, a fault $f_i \in F$ is considered to be isolable from another fault $f_j \in f$ if there exists a residual $R_k \in R$ that is sensitive to f_i but not to f_j. Due to the fact that the isolability is defined by pair of two faults, to isolate n_f faults, it is necessary to meet $\frac{n_f!}{(n_f-2)!}$ isolability requirements. This notion of isolability is thus more restrictive compared to the proposed definition 2.4.

The difference between the two notions of fault isolability

is illustrated on the following simple example. Consider a set of three residual generators and three faults with different sensitivities defined in Tab. 1. The symbol $*$ indicates that a given residual r_i is sensitive to a fault f_j.

$$R_{all} = \{r_1, r_2, r_3\} \quad F = \{f_1, f_2, f_3\}. \tag{3}$$

Table 1. Fault signature matrix.

	r_1	r_2	r_3
f_1	0	$*$	$*$
f_2	$*$	$*$	0
f_3	$*$	0	$*$

To isolate all three faults using the fault isolability requirements employed in previous works, the set of residuals should respect six different constraints:

$$\begin{aligned} &c_1 : f_1 \times f_2 \;\; c_2 : f_1 \times f_3 \;\; c_3 : f_2 \times f_1 \\ &c_4 : f_2 \times f_3 \;\; c_5 : f_3 \times f_1 \;\; c_6 : f_3 \times f_2 \end{aligned} \tag{4}$$

where $f_i \times f_j$ denotes a constraint that requires a residual sensitive to f_i but not to f_j.

Analysing the fault signature from Tab. 1, all three residuals are thus required to meet the six fault isolability constraints. However, it is possible to find smaller subsets of R_{all} capable of detecting and isolating all faults (3) using the fault signature isolability concept. The number of isolability constraints is then divided by two:

$$c_1 : S_1 \neq S_2 \;\; c_2 : S_1 \neq S_3 \;\; c_3 : S_2 \neq S_3. \tag{5}$$

It can be checked that any pair of residuals generates a unique fault signature for each fault, respecting the constraints 5, and is, therefore, a solution to the optimization problem (2).

4. Minimal Residual Selection Algorithm

A new algorithm to solve the optimization problem (2) is proposed. First, the minimal number of residuals needed to isolate n_f faults is calculated. Assuming that each residual r_i has only two states: $r_i = 0$ when $F_i = 0$ and $r_i \neq 0$ when $F_i \neq 0$, where F_i is the vector of faults that affects r_i. The lowest number of residuals n_{min} necessary to isolate n_f faults must follow the inequality:

$$2^{n_{min}} \geq n_f + 1. \tag{6}$$

It must be highlighted that n_{min} represents the theoretical lowest number of residuals necessary to generate n_f different fault signatures. The existence of such subset will depend on the sensitivity of each residual.

For instance, to isolate the three faults from Eq. (3), at least two residuals are required to generate three different fault signatures, considering that the fault signature $(0, 0)$ is excluded because it is equivalent to the fault-free state. For comparison, the solution proposed in (Jung & Frisk, 2018) based on optimization finds a subset of six residuals to isolate four faults.

The main idea behind the proposed algorithm consists in taking all possible combinations of n_{min} at a time of residual generators in R and checking if this subset of R generates a different fault signature for each residual. Assuming that the total number of residual generators is n_R, the number of all possible combinations is defined as

$$n_c = \frac{n_R!}{(n_R - n_{min})! n_{min}!} \tag{7}$$

From Eq. (7), if the number of residual generators is too big, it would be impossible to test the isolability properties of all possible subsets of R. For example, to isolate thirty faults ($n_f = 30$) using sixty residual generators ($n_R = 60$), it is necessary to have at least five residuals ($n_{min} = 5$), and there are more than five million possible combinations of residuals to be tested ($n_c > 5 \times 10^6$).

To restrict the number of residual generators, two new concepts are introduced:

- **Detectability class:** for each fault $f \in F$, list every residual sensitive to this fault $R_{df} \in R_{all}$;
- **Undetectability class:** for each fault $f \in F$, list every residual that is not sensitive to this fault $R_{uf} \in R_{all}$.

The idea is to select the most suitable residual generator for each detectability and undetectability class. The criterion for selecting that residual generator will depend on the residual generator method. In this work, the observer-based residual generation method is used. Two criteria are defined to choose the most suitable residual generator from a state-observer point of view:

1. Choose the residual generators composed of Ordinary Differential Equations (ODEs) or Differential Algebraic system of Equations (DAE) of index 1.
2. Select the residual generators with minimal "state cardinality," which means the residual with a minimal number of equations, which is equivalent to the state dimension of the corresponding observer.

The first criterion is related to the observer theory, which is mostly based on ODE systems. DAE systems of index one are also included because they can be easily transformed into an ODE by taking the derivative of the algebraic equations (Campbell, Linh, & Petzold, 2008).

The second criterion is related to model uncertainty. Each equation has a level of uncertainty due to modeling errors. It is thus suitable to choose the residuals with fewer equations to minimize the combined level of uncertainty.

Finally, the union of all residuals that meet both criteria for each detectability and undetectability class is used to test all possible combinations to verify the fault isolability requirements. The formal description of the process to find the minimal subset of residual generators is described in Algorithm 1. It can be divided into two main loops. The first loop takes the set of residual generator R and filters it using the two criteria defined above. The isolability properties of the filtered subset of residuals R_f are inspected. If the isolability properties are not met, a flag to relax the filtering constraints ($rCons$) is activated. The second loop tests all possible subsets of R_f based on the minimal number of residuals (n_{min}) needed. If no subset of R_f containing n_{min} residuals is capable of detecting and isolating the faults F, the minimum number of residuals n_{min} is increased, and the search restarts. The procedure returns a list R_{min} containing all subsets with n_{min} residuals that can detect and isolate all faults. The other procedures used in Algorithm 1 are described below.

- DETECTABILITYCLASS(R, F) for each fault $f \in F$, lists all residuals from R that are sensitive to this fault. Returns n_f subsets of residuals corresponding to each fault.
- UNDETECTABILITYCLASS(R, F) for each fault $f \in F$, lists all residuals from R that are not sensitive to this fault. Returns n_f subsets of residuals corresponding to each fault.
- FILTERRESIDUALS($d, u, rCons$) for each detectability class d and undetectability class u, filter the residuals considering cardinality and equations structure criteria. If the flag $rCons$ is activated, the cardinality criteria are relaxed. Returns the list of residuals R_f that fits all filtering criteria.
- CHECKISOLABILITY(R, F) checks if a group of residuals R generates unique fault signatures for each fault $f \in F$. Returns 1 if true and 0 if false.
- COMPUTESUBSETS(R, n_{min}) compute all possible combinations of residuals from R separated into groups of n_{min} residuals. Returns a list containing all possible combinations.

Algorithm 1 Residual Selection Algorithm

Inputs: Set of residual generators R, List of faults F
Output: Subsets of R with minimal cardinality R_{min}

```
procedure RESIDUALSELECTION(R,F)
    d ← DETECTABILITYCLASS(R,F)
    u ← UNDETECTABILITYCLASS(R,F)
    rCons ← 0
    isol ← 0
    while isol = 0 do
        R_f ← FILTERRESIDUALS(d, u, rCons)
        if CHECKISOLABILITY(R_f,F) then
            isol ← 1
        else
            rCons ← rCons + 1
    n_min ← COMPUTENUMMINRES(F)
    while R_min = ∅ do
        R_S ← COMPUTESUBSETS(R_f,n_min)
        k ← 0
        for all R_i ∈ R_S do
            if CHECKISOLABILITY(R_i, F) then
                R_min(k) ← R_i
                k ← k + 1
            if R_min = ∅ then
                n_min ← n_min + 1
    return R_min
```

5. RESIDUAL EVALUATION PROBLEM

The Algorithm 1 presented in Sec. 4 returns all possible combinations of residuals with minimal cardinality capable of isolating the predefined faults. In example (3), there are three different pairs or residuals that could be used to isolate the faults. A method to compare the residual generators is presented here. The objective is to quantitatively measure whether one residual is more sensitive than another.

Assuming that state observers will generate the residual by measuring the difference between the output estimated by the state observer $\hat{y}$ and the measured output y, the idea is to quantify the "innovation" that the fault will have on the measured states of the residual generator. Taking y_{bf} as the measured output before fault and y_{af} the measured output after fault, the innovation is defined as

$$In = y_{bf} - y_{af}. \tag{8}$$

In theory, the innovation brought by the fault is important to better the sensitivity of the residual generator to this fault. This procedure is illustrated on the same simple example used in (3). Consider a linear time-invariant system composed of a chain of integrators:

$$\begin{aligned} &e_1 : \dot{x_1} = k_1 x_2 \;\; e_2 : \dot{x_2} = k_2 x_3 \;\; e_3 : \dot{x_3} = k_3(u + f_3) \\ &e_4 : y_1 = x_1 + f_1 \;\; e_5 : y_2 = x_2 + f_2 \;\; e_6 : y_3 = x_3 \end{aligned} \tag{9}$$

where the unknown variables are $x = \{x_1, x_2, x_3\}^T$, the outputs are $y = \{y_1, y_2, y_3\}^T$, the input is u, the fault vector is $F = \{f_1, f_2, f_3\}^T$, and $k_i, i \in [1, 3]$ are known constants.

Thee residual generators candidates can be extracted from (9) using structural analysis, where three MSOs are computed and taken as residual generators. They are composed of the following equations

$$\begin{aligned} &r_1 = \{e_2, e_3, e_5\} \;\; r_2 = \{e_1, e_4, e_5, e_6\} \\ &r_3 = \{e_1, e_3, e_4, e_6\} \end{aligned} \tag{10}$$

the fault signature of the residuals (10) are illustrated in Tab. 1. It has been shown previously that any pair of (10) is enough to isolate the three faults. However, the impact of the faults is different for each residual.

For instance, let us compare the sensitivity of r_1 and r_2 when f_2 is injected using the procedure described above. For r_1, the relation between the fault f_2 and the output is direct because the output of r_1 is y_2. For r_2, the measurement of y_2 is first used to estimate $\dot{x}_1$, which is then used to estimate the output y_1. The innovation brought by f_2 in r_1 and r_2 can be summarized as

$$In_{r_1,f_2} = f_2 \quad In_{r_2,f_2} = k_1 f_2. \tag{11}$$

If the known constant k_1 is bigger than one, this empirical analysis indicates that r_2 will be more sensitive to r_1 to detect f_2. Repeating this analysis to the other faults and residuals, an efficient subset of residual generators concerning fault sensitivity can be found.

6. Application Example

The algorithm presented in Section 4 is used to find a list of minimal subsets of residuals capable of detecting and isolating a predefined list of faults in a multi-engine propulsion cluster of a reusable launcher.

6.1. Multi-Engine Propulsion Cluster Description

The propulsion cluster considered here is composed of three main parts: propellant tanks, feeding lines, and liquid-propellant rocket engines. The tanks are where the propellant is stored, and the feeding lines connect the propellant tanks with the rocket engines, where the thrust is generated. The propulsion cluster considered is composed of three rocket engines. The rocket engines use liquid oxygen (LOX) and liquid hydrogen H_2 as propellants (Pérez Roca, 2020). A simplified scheme of the LOX part of the multi-engine cluster is illustrated in Fig. 1. The feeding lines architecture with one main line splitting into secondary lines is optimal for minimum mass and pressure drop values (Miquel, 2020).

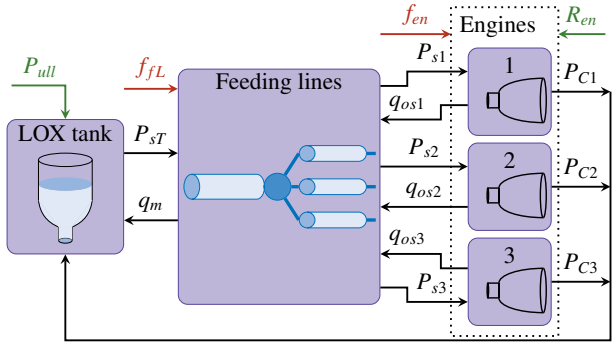


Figure 1. Multi-engine propulsion cluster scheme

The cluster operation can be summarized as follows: the rocket must follow a predefined trajectory. The trajectory is converted into thrust reference (R_{en}) and then given to the engines. Each engine has its control law that uses the control valves to meet the references. The valve position defines how much mass flow (q_{oi}) is used by the engine and, therefore, defines the mass flow that goes out of the tank through the feeding lines. The outlet pressure of the feeding lines (P_{si}) is imposed as the input pressure of the engines. P_{si} also depends on the tank's outlet pressure P_{sT}. The tank's outlet pressure is defined by the tank's ullage pressure (P_{ull}) and the rocket acceleration a_L. The acceleration depends on the thrust generated by each motor, which is directly related to the engines' combustion chamber pressure P_{Ci}. The index $\cdot_i \in \{1,2,3\}$ denotes the respective rocket engines.

The state vector is composed of the following variables

$$\begin{aligned} x = \{&P_{sT}, P_m, P_{si}, q_m, q_{si}, q_{GHi}, q_{CHi}, q_{THi}, q_{TOi}, \\ &P_{Ci}, P_{Gi}, P_{THi}, P_{TOi}, \omega_{Hi}, q_{GOi}, q_{COi}, \omega_{Oi}\}^T \end{aligned} \tag{12}$$

where P_{sT} is the output pressure of the LOX tank, P_m/q_m are the output pressure/mass flow of the main line, P_{si}/q_{si} are the output pressure/mass flow of the i-th secondary line, q_{GHi}/q_{GOi} are the gas generator H_2/LOX mass flow, q_{CHi}/q_{COi} are the combustion chamber H_2/LOX mass flow, q_{THi}/q_{TOi} are the turbine H_2/LOX mass flow, P_{Ci}/P_{Gi} are the combustion chamber/gas generator pressure, P_{THi}/P_{TOi} are the H_2/LOX turbine intake pressure, and ω_{Hi}/ω_{Oi} are the H_2/LOX pump rotating speed.

The vector of known variables z is composed of input variables u and output measurements y, $z = \{u, y\}$

$$\begin{aligned} u &= \{V_{GOi}, V_{GHi}, V_{COi}, V_{CHi}, V_{Zi}\} \\ y &= \{P_H, P_{ull}, P_m, P_{si}, P_{Ci}, P_{Gi}, \omega_{Hi}, \omega_{Oi}, \\ &\quad RMC_i, RMG_i\}^T \end{aligned} \tag{13}$$

where VG_{Hi}/V_{GOi} are the gas generator H_2/LOX control valves, V_{CHi}/V_{COi} are the combustion chamber H_2/LOX control valves, V_{Zi} is the valve that directs the gas generated by the gas generator to the turbines, P_H is the outlet pressure of the H_2 line, P_{ull} is the ullage pressure at the LOX tank, and RMC_i/RMG_i are the combustion chamber/gas generator mixture ratios.

The mixture ratios are the relation between the LOX and H_2 mass flows:

$$RMC_i = \frac{q_{COi}}{q_{CHi}} \quad RMG_i = \frac{q_{GOi}}{q_{GHi}}. \tag{14}$$

The fault vector f is composed of the following faults:

$$f = \{f_{qCOi},\ f_{VCHi},\ f_{VGHi},\ f_{\omega Hi},\ f_{RMCi}\}^T \tag{15}$$

where f_{qCOi} is a external LOX leakage in the combustion chamber, f_{VGHi} is a blockage in V_{GHi}, f_{VCHi} is a blockage in V_{CHi}, and $f_{\omega Hi}/f_{RMC}$ are bias faults in the sensors of ω_{Hi}/RMC_i respectively.

All the equations that describe the relations between states, inputs, and faults are listed in Appendix 7.

Considering that each engine is identical, only the measurements and faults from engine one are considered to simplify the implementation of the algorithm and avoid unnecessary computational. However, all results obtained for engine one can be automatically extended to the other two engines.

The first step is to find all residual generator candidates R. This step is performed using the Fault Diagnosis Toolbox (Frisk, Krysander, & Jung, 2017). In total, considering the equations and measurements of only engine one, the system is composed of fifty-three equations, and the degree of redundancy is eight. The residual generator candidates are obtained by computing all the Minimally Structurally Overdetermined (MSO) sets. Each MSO is a subsystem with a degree of redundancy one, i.e. it has one more equation than the number of unknown variables. All MSOs can be solved independently and are, therefore, residual generator candidates.

6.2. Algorithm implementation

The computation of all residual generators results in 24433 candidates that can possibly be used to detect and isolate five faults. From (6), at least three residuals are used to detect and isolate five faults.

The first loop of Algorithm 1 finds a subset of 16 residuals that meet both cardinality and state-observer criteria.

$$\begin{aligned} R_f = \{&r_{166}, r_{167}, r_{170}, r_{710}, r_{711}, r_{713}, r_{1000}, r_{1001}, \\ &r_{1006}, r_{1085}, r_{1320}, r_{1321}, r_{1326}, r_{1408}, r_{1593}, r_{1838}, \}. \end{aligned} \tag{16}$$

The fault sensitivity of the selected residuals is expressed in Table 2. It shows that the selected residuals 16 are enough to detect and isolate all faults considered.

The second loop of Algorithm 1 takes the selected residuals R_f from equation (16) and tests all possible combinations using the minimum number of residuals and selects the combination that generates unique fault signatures for each fault. In the first iteration of the loop, the minimum number of residual generators is three, and from (6), there are 560 possible combinations to be tested. However, there are no subsets of three residuals capable of isolating all faults. In the second iteration, the minimum number of residuals is increased by one, resulting in 1820 possible combinations of four residuals. The Algorithm 1 returns 40 combinations, each one containing four residuals from (16) that can isolate all faults. For comparison, the algorithm proposed in (Svärd et al., 2013) returns a subset of at least seven residuals to isolate the same five faults.

Two possible subsets of residuals are chosen for further analysis:

Table 2. Fault signature matrix.

	f_{VGH1}	f_{VCH1}	f_{qCOi}	$f_{\omega Hi}$	f_{RMCi}
r_{166}	0	0	0	$*$	0
r_{167}	0	0	$*$	$*$	$*$
r_{170}	0	0	$*$	$*$	$*$
r_{710}	$*$	$*$	$*$	$*$	$*$
r_{711}	$*$	$*$	0	0	0
r_{713}	$*$	$*$	0	0	$*$
r_{1000}	0	$*$	$*$	$*$	$*$
r_{1001}	0	$*$	$*$	$*$	$*$
r_{1006}	0	$*$	$*$	$*$	$*$
r_{1085}	0	$*$	$*$	$*$	$*$
r_{1320}	$*$	0	$*$	$*$	$*$
r_{1321}	$*$	0	$*$	$*$	$*$
r_{1326}	$*$	0	$*$	$*$	$*$
r_{1408}	$*$	$*$	$*$	$*$	$*$
r_{1593}	$*$	0	$*$	$*$	$*$
r_{1838}	$*$	$*$	$*$	$*$	$*$

$$\begin{aligned} R_1 &= \{r_{166}, r_{170}, r_{713}, r_{1001}\} \\ R_2 &= \{r_{166}, r_{170}, r_{713}, r_{1321}\} \end{aligned} \tag{17}$$

both subsets in (17) have almost the same structure; the only difference is the last residual generator. To compare those residual generators, the empirical residual evaluation method presented in Section 5 is used. The residuals are composed of the following variables:

- r_{1001}

$$\begin{aligned} x_{1001} &= \{q_{GO1}, q_{CH1}, \omega_{H1}\} \\ z_{1001} &= \{P_{s1}, P_H, P_{G1}, P_{C1}, RMC_1, RMG_1, \\ &\quad \omega_{O1}, \omega_{H1}, V_{CH1}, , V_{GO1}\} \end{aligned} \tag{18}$$

- r_{1321}

$$\begin{aligned} x_{1321} &= \{q_{GH1}, q_{CO1}, \omega_{H1}\} \\ z_{1321} &= \{P_{s1}, P_H, P_{G1}, P_{C1}, RMC_1, RMG_1, \\ &\quad \omega_{O1}, \omega_{H1}, V_{CO1}, , V_{GH1}\} \end{aligned} \tag{19}$$

the residuals have a very similar structure, having three states and the same output ω_H. One difference is when the fault f_{RMC1} is injected. In r_{1001}, RMC_1 is used to estimate q_{CO1}, on the other hand, in r_{1321}, RMC_1 is used to estimate q_{CH1}. When fault f_{RMC1} is injected, the estimation of the mass flows will be given by

$$\begin{aligned} r_{1001}: &\quad q_{CO1} = (RMC_1 + f_{RMC1}) q_{CH1} \\ r_{1321}: &\quad q_{HC1} = \frac{q_{CO1}}{RMC_1 + f_{RMC1}}. \end{aligned} \tag{20}$$

For residual r_{1321}, the influence of f_{RMC1} is directly observed in the output ω_{H1} because the evolution of ω_{H1} depends on q_{HC1}. Residual r_{1001} is not directly influenced be-

cause q_{CO1} is not used to estimate ω_{H1}. The fault f_{RMC1} will first impact the state q_{GO1} which then will influence the estimation of q_{GH1} and affects ω_{H1}. During those steps, the fault magnitude f_{RMC1} is divided by a constant bigger than one, attenuating the effect of the fault in the output. This makes the residual r_{1321} more suitable to detect f_{RMC1}. The same analysis can be extended to faults f_{qCO1} and $f_{\omega O1}$, where the magnitude of the faults is attenuated before affecting the output of residual r_{1001}.

6.3. Simulation results

To test the performance of the residuals (17) in simulation, an Unscented Kalman Filter (UKF) was calculated for each residual. This state estimator can deal with any type of nonlinearities and gives accurate estimations up to the third order of Taylor expansion (Wan & Van Der Merwe, 2000). The unscented transformation parameters were set at the default values, which gives an optimal solution for Gaussian distributions, with $\alpha = 0.001$, $\kappa = 0$, and $\mathcal{B} = 2$. The multi-engine cluster model was implemented using Simulink, and measurement noise was added. It is a white noise with zero mean, and the standard deviation varies according to the sensor specifications. For the rotational frequency of the turbopumps ω_{Oi}, ω_{Hi}, the Standard Deviation (SD) is 0.1% the rotational frequency when the engine at its nominal operating point. For the low-pressure values (P_{si}, P_H, P_{ull}), the SD is 0.1%, the nominal pressure value. For the high pressures (P_{Gi}, P_{Ci}) the SD is 0.2% the nominal value. The mixture ratios RMC_i and RMG_i have a standard deviation of 0.3%, the nominal value. The measurement noise covariance matrix $\mathcal{R}$ of the UKF was defined according to the standard deviations. The model parameters are considered perfectly modeled, so the process noise covariance matrix $\mathcal{Q}$ is proportionally defined ten times smaller than $\mathcal{R}$. To simulate the behavior of the system in closed-loop when a fault is injected, three PIDs were designed for each engine using the classical configuration (Pérez Roca, 2020). The PIDs use the valves to control the outputs $y_{PID} = [RMC_i, RMG_i, P_{Ci}]$. The closed loop system has a settling time to the step response of two seconds without overshooting.

Five faults are simulated in rocket engine 1, and each fault stays active for two seconds. The fault injection time and parameters are presented in Tab. 3.

The residuals generated by the UKFs are the difference between the output estimated by the state observer and the measured output, they are illustrated in Fig. 2. The UKF calculated from residual generator r_{166} is denoted UKF_{166}, etc.

From Fig. 2, the theoretical fault signature matrix of the residuals defined in Table 2 is observed in simulation. The exception is residual r_{1001} where the faults f_{qCO}, f_{wO} and f_{RMC} are attenuated by the residual's equations, and the impact of those faults cannot be seen when measurement noise is added. From simulation results, it is confirmed that the subset R_2 from (17) is more suitable for fault detection and isolation due to the higher sensitivity of r_{1321} when compared with r_{1001}.

Table 3. Fault parameters

Fault	Time	Effect
f_{VGH1}	12s-14s	Blockage 50% V_{GH1}
f_{VCH1}	16s-18s	Blockage 10% V_{CH1}
f_{qC1}	20s-22s	LOX Leak. of 0.8% nominal q_{CO1}
f_{wO1}	24s-26s	Bias of 1% nominal ω_{O1} value
f_{RMC1}	28s-30s	Bias of 5% nominal RMC_1 value

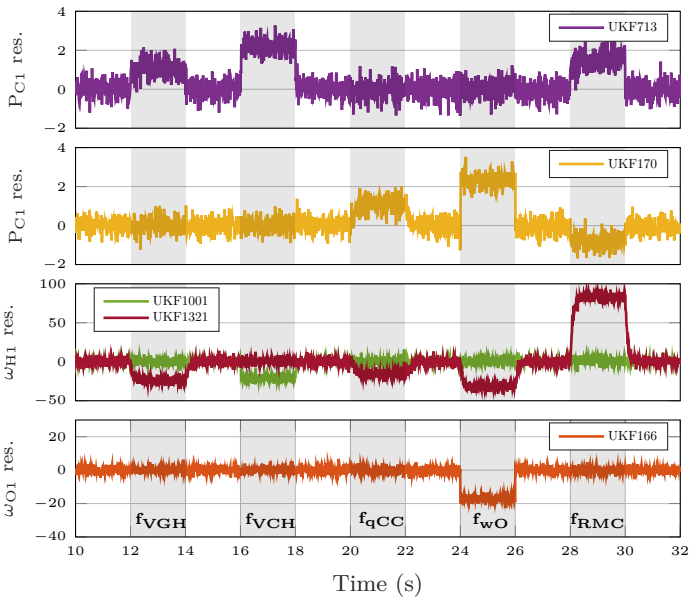


Figure 2. Residuals

7. Conclusion

A novel algorithm to find all possible subsets of residual generator candidates capable of detecting and isolating all faults with minimal cardinality has been presented. The minimal cardinality is achieved using a less restrictive isolability constraint based on the fault signature. Since the algorithm cannot be applied to a large number of residual generator candidates due to combinatorial explosion, two criteria to decrease the number of residual candidates were established. Those criteria take into account the residual generator method based on state observers, i.e. the reduced sensitivity to uncertainty when the number of state equations is minimal per residual. A procedure was presented to evaluate the selected residuals and compare the subsets with minimal cardinality returned by the algorithm. The proposed methods were applied in a model of a multi-engine propulsion cluster where five different faults

were considered. From 24433 residual generator candidates, the algorithm found 40 subsets, each one containing four different residual generators, that were capable of detecting and isolating the five faults. Two of those subsets of residual generators were implemented using Unscented Kalman Filter. Simulation results showed that the subsets can be used to detect and isolate all faults, and as a result the effectiveness of the proposed selection algorithm and quantitative sensitivity evaluation.

ACKNOWLEDGMENTS

This paper is a result of a study supervised by CNES and ONERA, involving experts in propulsion and FDI domains, to propose solutions for the diagnosis of reusable launchers.

REFERENCES

Campbell, S. L., Linh, V. H., & Petzold, L. R. (2008). Differential-algebraic equations. *Scholarpedia*, *3*(8), 2849.

Escobet, T., Bregon, A., Pulido, B., & Puig, V. (2019). *Fault diagnosis of dynamic systems*. Springer.

Frisk, E., Krysander, M., & Jung, D. (2017). A toolbox for analysis and design of model based diagnosis systems for large scale models. *IFAC-PapersOnLine*, *50*(1), 3287–3293.

Isermann, R. (2005). Model-based fault-detection and diagnosis–status and applications. *Annual Reviews in control*, *29*(1), 71–85.

Jung, D., & Frisk, E. (2018). Residual selection for fault detection and isolation using convex optimization. *Automatica*, *97*, 143–149.

Jung, D., & Sundström, C. (2017). A combined data-driven and model-based residual selection algorithm for fault detection and isolation. *IEEE Transactions on Control Systems Technology*, *27*(2), 616–630.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Krysander, M., Åslund, J., & Nyberg, M. (2007). An efficient algorithm for finding minimal overconstrained subsystems for model-based diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(1), 197–206.

Luenberger, D. G. (1964). Observing the state of a linear system. *IEEE transactions on military electronics*, *8*(2), 74–80.

Majumdar, A., & Steadman, T. (2001). Numerical modeling of pressurization of a propellant tank. *Journal of Propulsion and Power*, *17*(2), 385–390.

Miquel, V. (2020). *Propellant feeding system of a liquid rocket with multiple engines.* KTH Royal Institute of Technology.

Pérez Roca, S. (2020). *Model-based robust transient control of reusable liquid-propellant rocket engines* (Theses). Université Paris-Saclay.

Simani, S., Fantuzzi, C., Patton, R. J., Simani, S., Fantuzzi, C., & Patton, R. J. (2003). *Model-based fault diagnosis techniques*. Springer.

Svärd, C., Nyberg, M., & Frisk, E. (2013). Realizability constrained selection of residual generators for fault diagnosis with an automotive engine application. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *43*(6), 1354–1369.

Wan, E. A., & Van Der Merwe, R. (2000). The unscented kalman filter for nonlinear estimation. In *Proceedings of the ieee 2000 adaptive systems for signal processing, communications, and control symposium* (pp. 153–158).

Zhang, J., & Rizzoni, G. (2017). Selection of residual generators in structural analysis for fault diagnosis using a diagnosability index. In *2017 ieee conference on control technology and applications (ccta)* (pp. 1438–1445).

APPENDIX

The Liquid-Propellant Rocket Engine models are all derived from (Pérez Roca, 2020). All variables used in the equations that are not states (12), inputs and outputs (13) or faults (15) are known constants. Only the equations of the oxygen side are presented. The equations of the hydrogen side have the same structure. The only difference is the index $._O$ is replaced by $._H$ on the hydrogen side. The effect of the faults in the dynamic equations of the cluster is highlighted in red.

The pressure at the output of the oxygen turbopump P_{pOi} is given using manufacturer data:

$$P_{pOi} = \left(\frac{ap_O}{\rho_O} + R_{OGC}\right)(q_{COi} + q_{GOi})^2 + bp_O(q_{COi} + q_{GOi})\omega_{Oi} + cp_O\rho_O\omega_{Oi}^2. \quad (21)$$

The evolution of the oxygen mass flow that enters the combustion chamber q_{COi} and the gas generator q_{GOi} are derived from conservation of the momentum equation:

$$\dot{q}_{GOi} = \frac{1}{I_{GO}}\left[P_{si} + P_{pOi} - P_{Gi} - \left(\frac{1}{2\rho_O(V_{GOi} + f_{VGOi})^2} + R_{OG} + R_{OGC}\right) q_{GOi}^2\right] \quad (22)$$

q_{COi} has the same structure, where the subscript $._G$ is replaced by $._C$.

The hot gases mass flows are given by

$$\dot{q}_{THi} = \frac{1}{I_{TH}}\left(P_{Gi} - P_{THi}\right) - ky_{TH}R_{outG}\frac{T_G}{P_{Gi}}q_{THi}^2 \quad (23)$$

$$\dot{q}_{TOi} = \frac{1}{I_{TO}}(P_{Gi} - P_{TOi}) - Zr_iR_{outGG}\frac{T_G}{P_{Gi}}q_{TOi}^2 \quad (24)$$

where Zr_i is the equivalent resistive coefficient of the valve. The combustion chamber pressure P_{Ci} evolution can be approximated by first order Taylor expansion

$$\dot{P}_{Ci} = k1_C(q_{CHi} + q_{COi} - f_{qCOi}) - k2_C\sqrt{T_C}P_{Ci}. \quad (25)$$

The oxygen turbine pressure P_{TOi} is defined as

$$\dot{P}_{TOi} = k1_{TO}T_Gq_{TOi} - k2_{TO}\sqrt{T_G}P_{TOi}. \quad (26)$$

Finally, the rotational speed's evolution is given by manufacturer data

$$\dot{\omega}_{Oi} = \frac{1}{J_O}[T_{TOi} - \frac{ac_O}{\rho_O}(q_{COi} + q_{GOi})^2 - bc_O(q_{COi} + q_{GOi})\omega_{Oi} - cc_O\rho_O\omega_{Oi}^2] \quad (27)$$

where the motor torque T_{TOi} is given by $T_{TOi} = ST.W_i$ with ST the specific torque and W_i the work provided by the turbine pressure P_{TOi}.

For the feeding lines model, the evolution of the mass flow q and outlet pressure P in one rigid pipe, considering the effects of the fluid inertia, dynamic compressibility and neglecting the fluid thermal expansion, can be described by the momentum and mass balance equations:

$$\begin{aligned}\dot{q} &= \frac{S}{L}\left(P_{in} - P - \frac{f_rL}{2\rho S^2D}q^2\right)\\ \dot{P} &= \frac{\alpha^2}{V}(q - q_o)\end{aligned} \quad (28)$$

this pair of equations must be repeated for each pipe to model the feeding lines illustrated in Fig. 1.

The governing equations of pressurization of a propellant tank are obtained from (Majumdar & Steadman, 2001). The output pressure of the tank is defined as

$$P_{sT} = P_{ull} + \rho_O[a_L + g\cos(b)]H_d \quad (29)$$

Considering that a cylinder can approximate the shape of the tank, the gravitational head H_d is defined as

$$H_d = \frac{V_{LOX}}{\pi r^2}, \quad V_{LOX} = V_{LOX_0} - \frac{\int_0^t qm\,dt}{\rho_0} \quad (30)$$

The rocket's acceleration a_L can be approximated by a bivariate quadratic function total thrust T generated by the engines and the mass of the rocket m_R:

$$a_L = k_{1a} + k_{2a}T + k_{3a}m_R + k_{4a}Tm_R + k_{5a}m_R^2 \quad (31)$$

Finally, the mass of the rocket m_R can be calculated as:

$$m_R = m_{R_0} - \int qm\,dt - \int q_H\,dt \quad (32)$$

where $q_H = q_{CH} + q_{GH}$ is the total hydrogen mass flow used by the three engines, and m_{R_0} is the initial mass of the rocket.

Robust Remaining Useful Life Prediction Using Jacobian Feature Regression-Based Model Adaptation

Prasham Sheth[1] and Indranil Roychoudhury[1]

[1] *SLB, Menlo Park, California, 94025, USA*
psheth2@slb.com
iroychoudhury@slb.com

ABSTRACT

The accurate and robust prediction of remaining useful life (RUL) is critical for enabling the proactive mitigation of fault effects rather than reacting to them. For RUL prediction, one must model nominal and faulty system behaviors and how different faults progress over time. Complex data-driven machine learning (ML) models may capture both nominal and fault progression by updating the model parameters at different stages. As new data are observed, these model parameters can be updated to keep the system model always accurate. However, complete retraining of these models is both data- and computation-intensive and unsuitable for dynamic, fast-changing environments requiring quick recalibration. This calls for efficiently adapting the model to new operating conditions or the system's current state. One such efficient way to recalibrate model parameters to newly observed data using Jacobian feature regression (JFR) is presented in Forgione, Muni, Piga, and Gallieri (2023), where a recurrent neural network (RNN) models the current behavior of the dynamic system. Then, any subsequent deviation of observed measurements and the RNN model is attributed to an "unacceptable degradation of the nominal model performance." To update the RNN model, Forgione et al. (2023) propose augmenting the current model with additive correction terms learned by implementing JFR on observed "perturbed system" data. In this paper, we propose an automated *online* framework to adapt the model efficiently to always reflect the system's current state and use it for accurate RUL prediction and select JFR as one such adaptation technique. We extend the implementation of JFR-based model adaptation to hybrid models and demonstrate JFR to be more sustainable than the other retraining methods. Finally, we showcase the application of this approach to the oil and gas industry. A testbed that simulates a digital synthetic oilfield is used to show the effectiveness of this adaptation-based RUL prediction technique.

1. INTRODUCTION

1.1. Motivation

The accurate and robust prediction of remaining useful life (RUL) is one of the most critical functions of prognostics and health monitoring for assets. RUL is the time before a system can no longer function nominally. Knowing how much useful time a system has can help prevent having to react to such a failure and plan steps to mitigate its effects, e.g., planning maintenance and repair, thereby supporting the seamless operation of the facility.

Predicting RUL is a well-explored problem and can be implemented using (1) *model-based* or science-based approaches, (2) *data-driven* approaches that use available data but completely agnostic of scientific domain knowledge, or (3) *hybrid* approaches that combine scientific theory with available data. For accurate RUL prediction, a model must capture the nominal behavior of the system as well as the progressively degraded behavior. In model-based approaches, multiple models can be developed for nominal and degraded system behavior for different possible faults. A single, sufficiently complex data-driven or hybrid model may capture both nominal and degraded system behavior. One approach to building a single model is to periodically update or adapt the model to new system observations so as to reflect reality accurately. Many possible ways are available in the literature that enable this updating or adaptation of the model, such as retraining of machine learning (ML) models, recalibration of model parameters based on some initially collected field data, and so on. However, many of these approaches are computationally expensive, have intense data requirements, and are unsuitable for dynamic, fast-changing environments that need quick recalibration requirements.

1.2. Related Work and Their Challenges

Wang, Zhao, and Addepalli (2020) provide a well-structured summary of the increasing interest in using deep-learning approaches, such as autoencoders, deep belief networks, recur-

rent neural networks (RNNs), and convolutional neural networks for RUL prediction. The authors present conventional model-based as well as hybrid RUL prediction approaches and provide an excellent summary of the scope of the research development undertaken regarding RUL prediction.

To highlight a few works that focus on predicting RUL, Lei et al. (2016) propose an approach that contains two modules: (1) indicator construction, which fuses the information from multiple features and constructs a health indicator that they referred to as weighted minimum quantization error to correlate the machinery degradation, and (2) RUL prediction block that uses a particle-filtering-based algorithm. Ma and Mao (2021) propose a convolution-based long short-term memory (CLSTM) for predicting the RUL of bearings. By showcasing the ability of the CLSTM architecture to predict RUL effectively, they validate the effectiveness of using the spatial and temporal features. Y. Zhang, Xiong, He, and Liu (2017) showcase the use of an LSTM-RNN based method for predicting the RUL of lithium-ion batteries, highlighting the approach's capability to capture the long-term dependencies of capacity degradation in batteries.

A major challenge of the existing RUL prediction approaches is that they are system-state-dependent; i.e., their effectiveness in predicting RUL accurately is hinged on the assumption that the system model accurately reflects the current and future (degraded) system states. However, collecting data representing the system in all possible scenarios is impossible for many systems in real life, especially when these systems are never allowed to degrade sufficiently. The data collection process for such a wide range of possibilities is also extremely expensive. Therefore, the availability of training data for such models is a big bottleneck, and as soon as the system goes into a level of degradation that is not represented in the data, the chances of the model effectively predicting RUL degrades exponentially. In such scenarios, having a dynamic model that adapts based on the system's state becomes necessary. Researchers have focused on developing different approaches by using techniques such as transfer learning, domain adaptation, and modeling the degradation process. This does not represent the exhaustive list of techniques but just highlights the major approaches being explored.

Transfer learning is one technique for adapting the model to use the previous learnings to improve generalization about the new task. In the case of the RUL prediction, the task remains the same but differs as the underlying data distribution changes; hence, the model has to be "transferred" to this new set of data points that belong to the new distribution. The change in distribution, as mentioned earlier, could result from different operating conditions or different states of the system because of the degradation or upgrades to the equipment of the system. *Domain adaptation* falls within the umbrella of transfer learning, wherein the main focus is to help the model adapt from one or more sources of domains to a target domain. Ding, Ding, Zhao, Cao, and Jia (2022) introduce a multisource domain adaptation network for RUL prediction of bearing under varying conditions. Their domain adaptation strategy functions in two stages where the domain-specific distribution is integrated with regressor adaptation. A fusion of LSTM and domain adversarial neural networks (DANN) to extract temporal information from the time-series data and learn the domain-invariant features, thereby successfully addressing the challenge of distribution shifts in data domains resulting from the different states of the systems is proposed in da Costa, Akçay, Zhang, and Kaymak (2020). Si, Hu, Chen, and Wang (2011) present an approach to predict RUL using a Wiener process with a nonlinear and time-dependant drift coefficient. It, in particular, involves designing a state-space model and using Bayesian filtering to update the drifting function parameter. The method is significant because of its potential application in online prediction, which is one of the critical requirements for such an RUL prediction framework. With different dynamics of the underlying system, determining the frequency of the updates is a critical task. L. Liu, Guo, Liu, and Peng (2019) introduce a data-driven framework for RUL prediction that integrates sensory anomaly detection and data recovery and improve RUL prediction by detecting sensory anomalies, recovering data, and using this recovered data for more accurate predictions. Huang, Xu, Wang, and Sun (2015) focus on addressing nonlinear degradation trajectories and heterogeneity in practical systems. They combine a nonlinear Wiener process with an adaptive drift feature. Y. Zhang, Yang, Xiu, Li, and Liu (2021) present an integrated technique that combines the Wiener process for degradation modeling and an LSTM network for forecasting degradation increments by learning the long-term dependencies of the offline degradation model and online observed degradation. Cheng et al. (2023) focus on predicting the RUL of the machinery under varying working conditions. Their proposed approach uses dynamic domain adaptation by integrating dynamic distribution and adversarial adaptation networks to predict RUL effectively. Pan, Li, and Wang (2022) propose combining LSTM and particle filter to predict the RUL for lithium-ion batteries under different stress conditions. They particularly leveraged transfer learning to update the LSTM model to ensure generalizability and then particle filter to capture the uncertainty. Siahpour, Li, and Lee (2022) introduce consistency-based regularization into the DANN training process. Consistency-based regularization helps to remove the negative impact of missing information. Sun et al. (2019) present a deep transfer learning network based on spare autoencoder. They incorporate three strategies — weight transfer, feature transfer learning, and weight update — to improve adaptability and prediction accuracy. Also, they showcased the network's capability to be trained on one tool and then being transferred to another tool under operation for online RUL prediction. The use of

bi-directional LSTM (BLSTM) neural networks and transfer learning for RUL estimation is explored in A. Zhang et al. (2018). They address the challenges of insufficient failure progression samples in data-driven prognostics by training the model on different but related datasets and then fine-tuning it with the real target domain dataset. J. Liu, Saxena, Goebel, Saha, and Wang (2010) present an adaptive recurrent neural network (ARNN) model for predicting the RUL of lithium-ion batteries. The model employs a dynamic state forecasting approach using a neural network architecture that adapts by optimizing the model's weights using the recursive Levenberg-Marquardt method.

1.3. Proposed Solution and Contributions

A better way to recalibrate model parameters to match model predictions to the newly observed data using Jacobian feature regression (JFR) is presented in Forgione et al. (2023). An RNN is used to model the dynamic system using available measurements. Then, as the system dynamics change, it causes the nominal model to be inaccurate for predicting the observed measurements in the presence of the perturbed system dynamics. The core idea of their approach is to adapt an existing RNN model, which was trained on data from a nominal system, to perturbed system dynamics, not by retraining the model from scratch, but by including an additive correction term to the nominal model's output. This correction term is designed to account for the discrepancies between the nominal system and the perturbed system. In other words, as an "unacceptable degradation of the nominal model performance" occurs, Forgione et al. (2023) propose a transfer learning approach to improve the performance of the nominal model in the presence of perturbed system dynamics, where the nominal model is augmented with additive correction terms that are trained on observed perturbed system data. These correction terms are learned through JFR "defined in terms of the features spanned by the model's Jacobian concerning its nominal parameters." Efficient model adaptation is achieved by using the JFR in the feature space defined by the Jacobian of the model with respect to its nominal parameters. Forgione et al. (2023) also propose a non-parametric view that uses the Gaussian process. This could be useful to provide flexibility and efficiency for very large networks or when only a few data points are available.

The contributions of this work are significant because they offer a more efficient and effective way to keep data-driven and hybrid models accurate when applied to dynamical systems that experience changes over time. We address some of the challenges described in Section 1.2 by building upon the method introduced in Forgione et al. (2023) as follows:

1. We present an automated approach to use adaptation techniques for predicting RUL while ensuring system-state-dependency of the models. Although JFR is used as the model-adaptation technique in this paper, JFR can be replaced by any other model-adaptation algorithm without any loss of generalizability.
2. We extend the implementation of JFR-based model adaptation to hybrid models that combine physics and data-driven models. This is important (and even necessary) as the representation of systems using data-driven models can become a bottleneck if the training data are limited, and there could be no guarantee that the data-driven models follows the physics of the system behaviors in all possible scenarios.
3. We highlight the lower carbon footprint of the JFR-based adaptation technique instead of retraining the model completely using the standard transfer learning.
4. We modify the offline adaptation approach into an online adaptation approach, which becomes critical to the PHM systems, especially in RUL prediction. To enable online adaptation, we use the anomaly detection output in order to trigger the model-adaptation.
5. Finally, we also discuss the application of our JFR-based model adaptation approach to assets relevant to the oil and gas industry. In particular, we discuss the results of applying the technique to a testbed that simulates a digital synthetic oilfield.

1.4. Organization

The remainder of this paper is organized as follows. Section 2 formulates the RUL prediction problem, and our approach to solve this. Section 3 includes the experimental setup and results, and finally, Section 4 concludes the paper and provides directions for future work.

2. Problem Setup and Approach

To set up the problem, let us denote a system by S, that typically takes in some inputs from discrete timesteps 1 to k, i.e., $\mathbf{x}_{1:k}$, and has measured outputs $\mathbf{y}_{1:k}$. Let us assume that M denotes a model representing this system S that takes in the same inputs $\mathbf{x}_{1:k}$ and outputs simulated measurements denoted by $\hat{\mathbf{y}}_{1:k}$. Due to differences between a model and reality, such as modeling error, and measurement noise, $\mathbf{y}_{1:k}$ and $\hat{\mathbf{y}}_{1:k}$ are seldom exactly the same, but a "good" model M would generate $\hat{\mathbf{y}}_{1:k}$ that is very close to the real outputs $\mathbf{y}_{1:k}$. We define a *threshold function* $T : \hat{\mathbf{y}}_\kappa \rightarrow \{\texttt{true}, \texttt{false}\}$ that partitions the operational state of the system into nonfailure and failure states based on observed measurements, such that $T(\mathbf{y}_\kappa)$ returns `true` when the system is in a failure state, and `false` otherwise. If the time of prediction is denoted by k_P, then typically, we define end-of-life (EOL) predicted at time k_p as $EOL(k_P) = \inf\{k' : k' \geq k_P \text{ and } T(\hat{\mathbf{y}}_k)\}$, and RUL at time k_P is defined as $RUL(k_P) = EOL(k_P) - k_P$. To make EOL and RUL predictions, the model M is fed hypothesized future inputs $\mathbf{x}_{k_P:\infty}$, also denoted by $\mathbf{x}_{future}$.

The model M can be defined/trained using model-based, data-

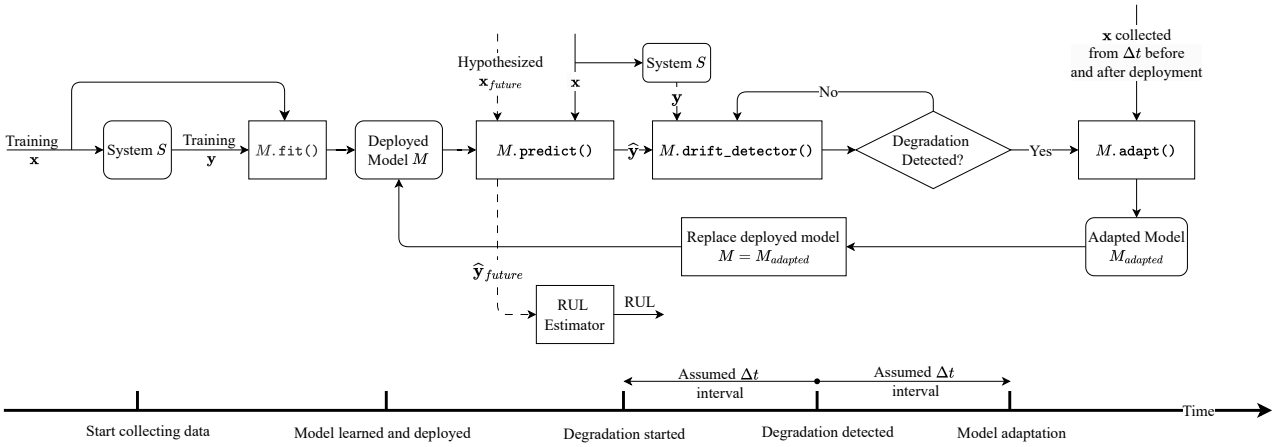


Figure 1. The workflow describing how the adaptation technique could adapt the model trained in the controlled setting (nominal model) to predict RUL after the degradation in the system is detected.

driven, or hybrid (model-based + data-driven) approaches. We use standard techniques to train or build such a model M and train it to the input-output signals using the $M.\texttt{fit()}$ algorithm. Typically nominal $\mathbf{x}$ and $\mathbf{y}$ combinations can be used to build M. Once the model is trained, it is deployed, and the $M.\texttt{predict()}$ algorithm passes new $\mathbf{x}$ through M to generate new predicted $\mathbf{y}$. $M.\texttt{predict()}$, when fed with hypothesized future inputs $\hat{\mathbf{x}}_{future}$, can predict estimated future outputs $\hat{\mathbf{y}}_{future}$ which is then fed to the RUL Estimator which passes $\hat{\mathbf{y}}_{future}$ through a threshold function T to compute the RUL of the system.

Now, ideally, if there is no degradation in the system, the M model would forever be able to correctly predict the future observations of the system. However, this is never the case as all engineered systems eventually encounter some sort of degradation or failure. Also, there is no guarantee that the operating conditions will remain constant throughout the system's life. One way to adapt to this changing system dynamics would be to retrain the model for every new pair of $\mathbf{x}$ and $\mathbf{y}$; however, that process is computationally wasteful. Hence, to intelligently call the model update, we develop and deploy an $M.\texttt{drift\_detector()}$ algorithm that compares the predictions from M, and the sensors obtained from the real system to see if there is a statistically significant drift between the predicted and observed sensors. If yes, then while there are many reasons for which this drift could occur, we attribute this drift to degradations in the system that are not captured by the deployed model M anymore, and the parameters of this model need to be re-calibrated or adapted to the newly observed sensors. If that is the case, then we call the $M.\texttt{adapt()}$ that helps us adapt the model to the new dynamics using the newly observed data. We denote this adapted model as $M_{adapted}$. In our case, the JFR-based model adaptation algorithm is used to adapt the model to new data observed. $M_{adapted}$ now replaces the model M, and the process continues until a significant deviation is *again* detected in the sensor readings predicted by M and the observed sensor readings from the system. Figure 1 presents an overview of the overall workflow.

As established, the data from the controlled environments could be used to configure and train different models that represent the system. Over time, as the system's behavior changes, the model becomes stale and its predictions are inconsistent with the system's behavior. As the system operates in real life, different measurements are collected and stored in some database. Using the designed approach, there are two choices for model adaptation.

Condition-Based Model Adaptation (CBMA): The data are continuously collected from the deployment environment in this setting. The model and the system are constantly monitored, and any kind of deviations are detected and tagged. If the deviation is above the threshold, all the past information collected before the deviation happens is used for adapting the model. It is an offline adaptation technique as not all the incoming information is directly used for adaptation. Rather the adaptation is triggered based on the output of anomaly detection. From an implementation perspective, for condition-based model adaptation, we use data from a window comprising of Δt time steps *before and after* the time at which degradation was detected, where Δt is a design choice. The timeline at the bottom of Figure 1 visually depicts this.

Continuous Model Adaptation (CMA): In this scheme of adaptation, there is no dependency on the anomaly detection process. As and when a new measurement is recorded it is used for adapting the model. This helps in the continuous utilization of the incoming information.

While both CBMA and CMA methods enable the efficient use of the incoming data to update the model continuously,

these two model adaptation methods have their pros and cons. Specifically, CMA requires high computing power as the models are continuously adapted. Furthermore, observing a single outlier can result in a deviation from the model's behavior, whereas it is not a persistent thing for the physical system. On the other hand, CBMA requires dependencies on the anomaly detectors and the storage systems where the data needed for model adaptation are stored. CBMA also enables us to quantify the model's behavior change which could be reflected based on the differences between the predictions of the nominal model that is trained using the data from a controlled setting, and the predictions of the adapted model that is adapted using incoming data predictions.

Our proposed adaptation approach can be used for either CMA or CBMA. Further, we have extended the offline model adaptation approach presented in Forgione et al. (2023) to hybrid models that represent the dynamics of a system by leveraging both first-principles domain knowledge and data-driven ML approaches. Karpatne, Watkins, Read, and Kumar (2017) presents physics-guided neural networks (PGNN), one such example of a hybrid modeling approach. PGNN uses the first principles model parallel to the data-driven components (e.g., RNNs). It could be helpful to directly use the first principle model even if it is not tuned and calibrated to the best quality. Such developed hybrid models for the specific systems could be further coupled with the adaptation technique to help us have a model that is always in close alignment with the physical system. A model that is always closely aligned with the physical system enables seamless deployment of different applications such as optimization, control, forecasting, prognostics and health management, automation, and decision-making, among others.

3. Experimental Setup and Results

Digital Synthetic Oilfield Testbed Setup: Our testbed's design is particularly chosen to mimic real-life oilfields. The test bed has three DC motor pumps attached to three flowmeters. Each DC motor pump pumps the water (used in place of oil) from the well to the eventual storage. The flowmeters measure the flow exiting the pumps. We also attached a fourth flowmeter to calculate the aggregated flow from the three pumps. Single and persistent faults are injected into each of the pumps to represent the loss of efficiency. The EOL condition for each pump is defined as the state when any pump's output flow dips below 0.15 units. The controllable input in the case of each pump is the pump speed, and the output measurement from the flowmeter would be the flow rate. Since the input voltage determines the pump speed, the voltage is considered the equivalent input variable for each of the pumps. Figure 2 represents the internal structure of the DC motor pump[1].

[1] https://ctms.engin.umich.edu/CTMS/?example=MotorSpeed\&section=SystemModeling

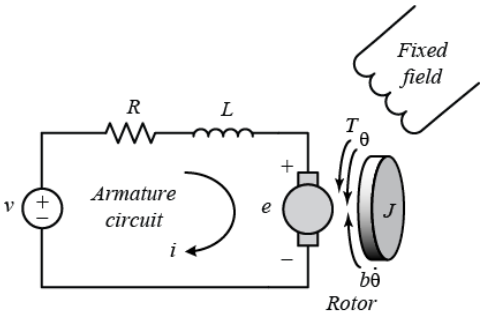


Figure 2. The electric equivalent circuit of a DC motor.

Based on the internal structure of each pump, the state-space model was designed for the testbed, considering the rotational speed and electric current as the state variables. After defining the established conditions to represent the system parameters, this state-space model for the testbed was used to simulate the operation of the three pumps in the oilfield. Equations (1a–1f) summarize the representation of this digital synthetic oilfield testbed. In this setup, (V_p for pump $p \in [1, 2, 3]$) represents the voltage and hence the controlled pump speed for each pump respectively (controllable inputs); ($y_p = \omega_p$, $p \in [1, 2, 3]$) represents the flow rate of each pump respectively (system measurements); and the hidden state variables include (ω_p, $p \in [1, 2, 3]$) that represents the angular momentum of each pump respectively, (i_p, $p \in [1, 2, 3]$) that represents the current drawn for each pump respectively. The inductance L_p, resistance R_p, and back electromotive force constant k_p for pump $p \in [1, 2, 3]$ are the system parameters.

$$\frac{d\omega_1}{dt} = \frac{1}{L_1}(V_1 - R_1 i_1 - k_1 \omega_1) \tag{1a}$$

$$\frac{di_1}{dt} = \frac{1}{J_1}(k_1 i_1 - B_1 \omega_1) \tag{1b}$$

$$\frac{d\omega_2}{dt} = \frac{1}{L_2}(V_2 - R_2 i_2 - k_2 \omega_2) \tag{1c}$$

$$\frac{di_2}{dt} = \frac{1}{J_2}(k_2 i_2 - B_2 \omega_2) \tag{1d}$$

$$\frac{d\omega_3}{dt} = \frac{1}{L_3}(V_3 - R_3 i_3 - k_3 \omega_3) \tag{1e}$$

$$\frac{di_3}{dt} = \frac{1}{J_3}(k_3 i_3 - B_3 \omega_3) \tag{1f}$$

There are multiple ways to model such systems. Neural state-space formulation is one such approach that could be used to model the system. This approach uses a couple of neural networks (NNs) to model state-transition and state-observation models. The same could be represented using Equations (2a–2b) where the function f represents the state-transition model and function g represents the state-observation model. In this neural state space formulation, once the model is trained, we

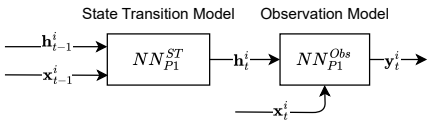


Figure 3. Schematic representation of neural state space model to model each component independently using just the inputs and hidden variable that affect it.

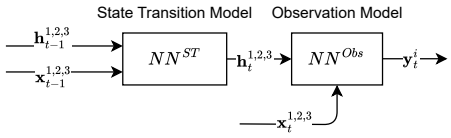


Figure 4. Schematic representation of neural state space model to model each component independently using all the inputs and hidden variables from the entire system.

utilize the forward Euler method to walk forward and make a closed-loop prediction for the next steps. Also note we are denoting the controllable inputs to the system by $\mathbf{x}$, the hidden state variables by $\mathbf{h}$, and the observed system measurements by $\mathbf{y}$. The subscript t represents the time step to which these values correspond.

$$\mathbf{h}_t = f(\mathbf{x}_{t-1}, \mathbf{h}_{t-1}) \tag{2a}$$
$$\mathbf{y}_t = g(\mathbf{x}_t, \mathbf{h}_t) \tag{2b}$$

Based on the described system for the testbed, we have a scenario where we have three pumps as the system's core components. For each of these three pumps, we have one controllable input and one measurement (output), and then internally, we have two state variables. Three major combinations were used for these different variables in our experiments. The summary and the description of why the particular selection was considered are described below.

1. One controllable input, one measurement, and three state variables: This setting enables us to represent each pump independently. Figure 3 represents this setting.
2. Three controllable inputs, one measurement, and seven state variables: In this setting, we model each pump's measurement independently but still consider all the input and underlying state variables. Figure 4 represents this setting. This enables us to consider all system dynamics together and learn the dependence of each pump's output on the entire system.
3. Three controllable inputs and three measurements: As shown in Figure 5, in this setting, we model the entire system using a single model that considers all the input variables and tries to learn the entire system by itself.

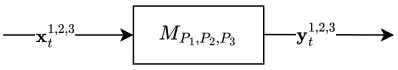


Figure 5. Schematic representation of the model when the entire system is modeled using the inputs and measurements.

In the first two settings, since we are able to use the available information for the state variables, the neural state space approach is used to model the system. There are existing approaches (e.g., Sheth, Roychoudhury, Chatar, and Celaya (2022)) that model the state-transition and state-observation function separately using two independent networks and in a joint setting where two networks are connected to each other. In this setting, these functions are represented using a NN. We also consider the measurement to be an unknown hidden state variable that needs to be estimated. In such a functional way, the state observation model becomes a passthrough function to select the state variable representing the measurement. Hence, the state-observation function could be omitted as represented in Equation 3. This particular method helps us condition the model in such a way that it has to produce the correct combination of the state variables as well as the measurement. By penalizing the wrong predictions, the model is regularized to adhere to the internal relations between state variables and the measurements. This could also be thought of as the state variables providing the regularization to the original model that is penalized for any sort of inconsistencies between the state variables and the measurement.

$$\mathbf{h}_t \cup \mathbf{y}_t = NN(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{y}_{t-1}) \tag{3}$$

The first two settings differ in the variables used by the NN to represent the state-space formulation. Equations 4 and 5 represent the two settings, respectively. The superscripted i represents the pump number to which different values correspond.

$$\mathbf{h}^i_t \cup \mathbf{y}^i_t = NN(\mathbf{x}^i_t, \mathbf{h}^i_{t-1}, \mathbf{y}^i_{t-1}); \quad i \in \{1,2,3\} \tag{4}$$

$$\mathbf{h}^i_t \cup \mathbf{y}^i_t = NN(\mathbf{x}^1_t, \mathbf{x}^2_t, \mathbf{x}^3_t, \mathbf{h}^1_{t-1}, \mathbf{h}^2_{t-1}, \mathbf{h}^3_{t-1}, \mathbf{y}^i_{t-1}); \quad i \in \{1,2,3\} \tag{5}$$

In the third setting, since the system is modeled as a whole, we utilize an LSTM network to model the system, and all the time dependence is captured through the LSTM network and can be represented as shown in Equation 6.

$$\mathbf{y}^1_t, \mathbf{y}^2_t, \mathbf{y}^3_t = LSTM(\mathbf{x}^1_t, \mathbf{x}^2_t, \mathbf{x}^3_t) \tag{6}$$

For all three settings, online and offline adaptation techniques have been integrated to adapt the model. Since the faults are injected independently into the pumps of the testbed, the summary table for the experimental results has four rows, one for the nominal setting and the remaining three for settings cor-

Table 1. Summary of experimental results (data-driven models). The metrics for all three pumps are shown separately under the columns for RMSE and R^2.

Scenario Number	Fault Setting	Number of Models	RMSE before Adaptation	RMSE after Adaptation	R^2 before Adaptation	R^2 after Adaptation
1	Nominal	3	[0.01, 0.01, 0.05]	-	[0.99, 0.99, 0.50]	-
	Fault in Pump 1	3	[0.06, 0.01, 0.05]	[0.01, 0.00, 0.01]	[-0.54, 0.99, 0.50]	[0.98, 0.99, 0.99]
	Fault in Pump 2	3	[0.01, 0.06, 0.05]	[0.00, 0.01, 0.01]	[0.99, -0.60, 0.50]	[0.99, 0.98, 0.99]
	Fault in Pump 3	3	[0.01, 0.01, 0.02]	[0.00, 0.00, 0.00]	[0.99, 0.99, 0.78]	[0.99, 0.99, 0.99]
2	Nominal	3	[0.01, 0.01, 0.02]	-	[0.98, 0.96, 0.92]	-
	Fault in Pump 1	3	[0.07, 0.01, 0.02]	[0.00, 0.00, 0.01]	[-0.98, 0.96, 0.92]	[0.99, 0.99, 0.98]
	Fault in Pump 2	3	[0.01, 0.07, 0.02]	[0.00, 0.01, 0.01]	[0.98, -1.22, 0.92]	[0.99, 0.99, 0.98]
	Fault in Pump 3	3	[0.01, 0.01, 0.05]	[0.00, 0.00, 0.01]	[0.99, 0.99, -0.04]	[0.99, 0.99, 0.98]
3	Nominal	1	[0.01, 0.02, 0.04]	-	[0.96, 0.95, 0.7]	-
	Fault in Pump 1	1	[0.06, 0.02, 0.04]	[0.01, 0.01, 0.01]	[-0.49, 0.95, 0.7]	[0.96, 0.99, 0.97]
	Fault in Pump 2	1	[0.01, 0.05, 0.01]	[0.01, 0.01, 0.01]	[0.96, -0.14, 0.7]	[0.99, 0.96, 0.97]
	Fault in Pump 3	1	[0.01, 0.02, 0.03]	[0.01, 0.01, 0.01]	[0.97, 0.95, 0.53]	[0.99, 0.99, 0.95]

Table 2. Summary of performance of physics-based model. The metrics for all three pumps are shown separately under the columns for RMSE and R^2.

System Setting	RMSE	R^2
Nominal	[0.04, 0.04, 0.01]	[0.81, 0.83, 0.99]

responding to the faults in three pumps, respectively.

Table 1 summarizes the results from the experiments to model the system, where each model's performance is evaluated using the root mean square error (RMSE) and R^2 metrics, where

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}},$$

and

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}.$$

For the sake of simplicity, the aggregated results from the online experiments are shown. Figure 6 shows the plots depicting the nominal model's prediction, predictions from the adapted model, and the actual system behavior when the fault was present in Pump 1. It also shows the threshold value, which could be used to predict RUL. Essentially, the time when the flow value for any pump goes below the threshold could be considered as the RUL for the system.

(Karpatne et al., 2017) introduced PGNN that enables us to couple the physics-based model with an NN. The core idea behind the coupling is to allow the NN to overcome the regions where the physics-based model might make errors because of the lack of generalizability introduced by the poorly estimated parameters. (Sheth et al., 2022) have successfully demonstrated the advantages of integrating the PGNN with neural state-space models. Figure 7 represents the original structure of the PGNN and Figure 8 represents the modified structure of PGNN for the neural state-space model as designed in (Sheth et al., 2022)

Inspired by these works and the ensuring need for the generalizability and scientific accuracy of the models representing the system, we have implemented all three scenarios using hybrid models that combine physics-based models with data-driven models. Since we designed the testbed, we can access the actual physics model used to collect the data. However, to mimic the scenarios we have in real life where the exact physics model is unavailable, we decided to use the functional form of the original physics model. We estimated the parameters after introducing some random noise to the recorded measurements. In doing so, we estimated the parameters of the physics model, which were not completely aligned with the underlying system model. This mimics the scenario of having a physics-based model that is not well-calibrated. Table 2 summarizes the performance of the physics-based model in the nominal setting.

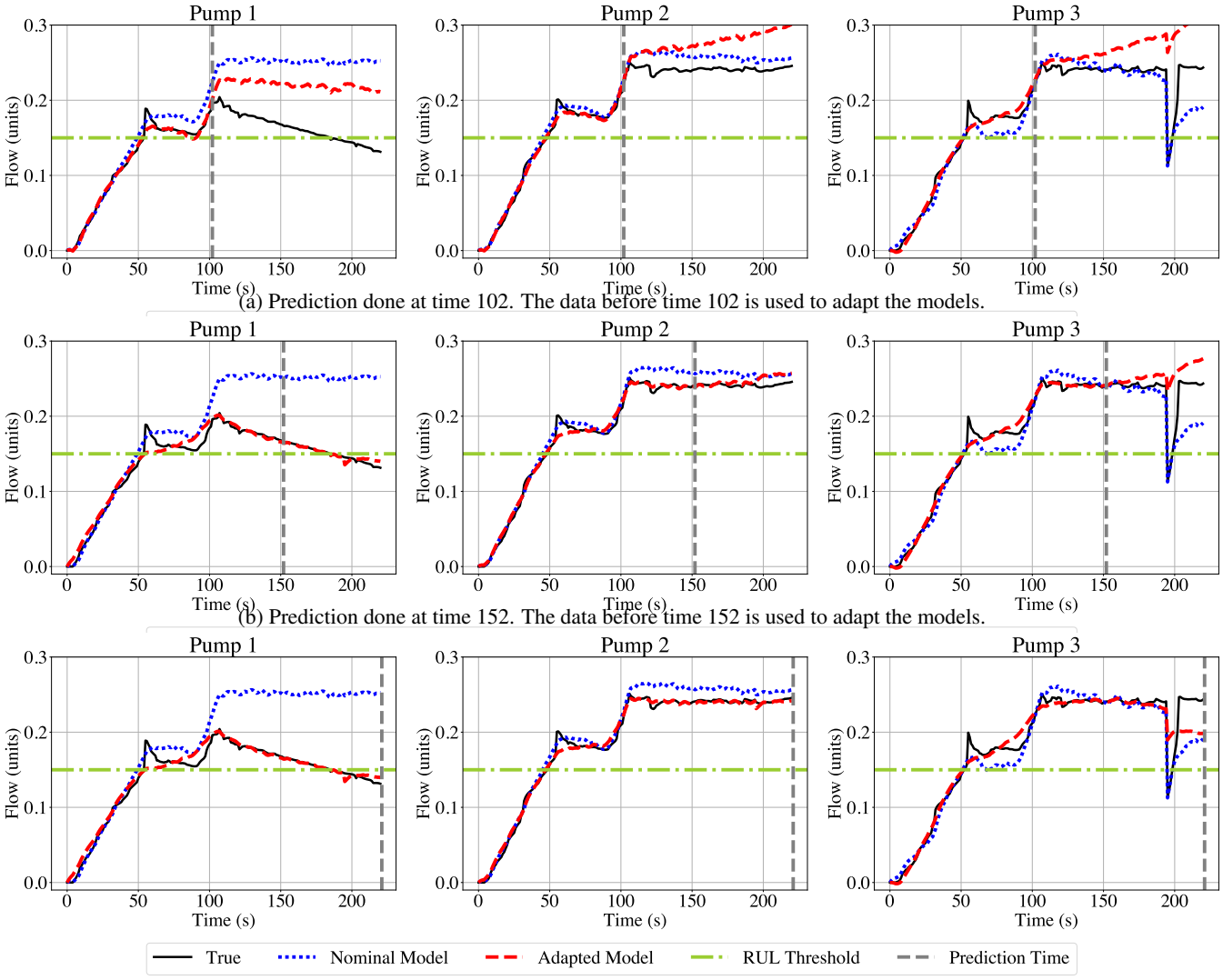
Based on the description of our neural state-space model and the PGNN architecture, there are two major ways in which the output from the physics model could be used:

1. Using the output estimate from the physics model as an input to the data-driven model. This strictly represents the PGNN architecture described in Figure 7. Figure 9 represents the same in our scenario.

2. Treating the output estimate from the physics model as one of the state-transition variables in the neural state-space formulation. This way, the output estimate from the physics model is integrated into the state-transition part of the neural state-space model. The output estimate from the physics model from the previous timestep is considered while predicting the actual output for the current timestep. Also, the model is penalized for pre-

Table 3. Summary of experimental results (physics-regularized data-driven models). The metrics for all three pumps are shown separately under the columns for RMSE and R^2.

System Setting	RMSE	RMSE after Adaptation	R^2	R^2 after adaptation
Nominal	[0.01, 0.02, 0.02]	-	[0.99, 0.94, 0.90]	-
Fault in Pump 1	[0.06, 0.02, 0.02]	[0.01, 0.01, 0.01]	[-0.71, 0.94, 0.90]	[0.98, 0.99, 0.99]
Fault in Pump 2	[0.01, 0.08, 0.02]	[0.01, 0.01, 0.01]	[0.99, -1.46, 0.90]	[0.99, 0.99, 0.99]
Fault in Pump 3	[0.01, 0.02, 0.06]	[0.01, 0.01, 0.02]	[0.99, 0.94, 0.38]	[0.99, 0.99, 0.90]



(a) Prediction done at time 102. The data before time 102 is used to adapt the models.

(b) Prediction done at time 152. The data before time 152 is used to adapt the models.

(c) Prediction done at time 221 (End of the simulation). The data before time 221 is used to adapt the models.

Figure 6. Flow estimates for the three pumps of the system representing the oilfield using the data from different timesteps to adapt the model. The blue dotted line represents the estimates from the nominal model, the red dashed line represents the estimates from the model after adaptation, and the black solid line represents the actual system behavior when the fault has been introduced in Pump 1. The green dashdot horizontal line represents the threshold that could be used to determine the RUL of the system. The vertical gray dashed line represents the present time till which the data from the system are observed.

dicting the wrong value for the output estimate from the physics model. Thus, the output estimate for the physics model works both as a signal for predicting the system's output from the model and provides a regularization effect for the model to be grounded to the physical relationships captured by the physics model. We refer to this set-

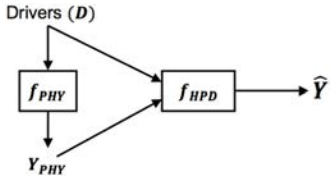


Figure 7. Structure of PGNN (Karpatne et al., 2017).

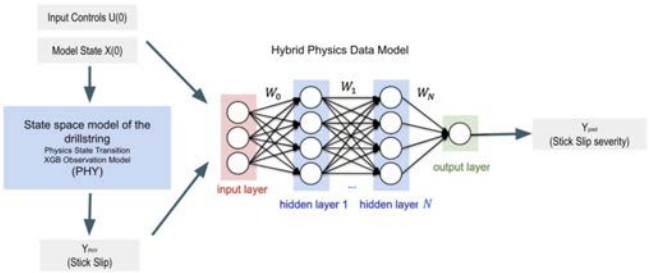


Figure 8. Schematic representation of PGNN for neural state-space model as shown in (Sheth et al., 2022).

ting as physics-regularized neural network (PRNN) and is shown in Figure 10. The predicted physics model estimate is compared with the actual physics model estimate. It is added to the loss function for training, and in the inference phase, we can ignore this output.

In our experiments, we evaluated both techniques and we coupled the physics model with the neural state-space models that we have for the second scenario. Table 3 summarizes the results from the experiments for the PRNN model corresponding to the second setup where the output of the physics model is used as a regularization condition. Figure 11 showcases the prediction estimates when the nominal version of the PRNN model is used for forward Euler simulation to generate the predictions for all timesteps in the closed loop setting and the prediction estimate for the same system state with the adapted PRNN model. Conducting experiments with PRNNs in this particular setting validates how the adaptation technique has been integrated into the neural state-space model, thus enabling the adaptation of hybrid models representing the system. Similar to the neural state-space model, the estimates of the physics-based model could be integrated into the list of controllable inputs for the LSTM model. When experiments for this particular setting were conducted, similar to the neural state-space model, positive results were obtained.

Based on the presented results in Table 1 and Figure 6, it is evident that the adaptation technique helps robustly adapt to the fault scenario. Further, it helps slightly reduce the errors due to the noise in measurements for the components without any faults, thus improving the overall prediction. Comparing the performance of the nominal model learned using the physics with the PRNN (first row of Table 2 and Table 3), it could be seen that the PRNN helps improve the performance over

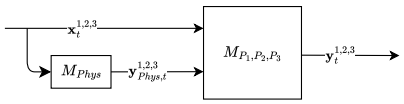


Figure 9. Schematic representation of PGNN for neural state-space model for our testbed.

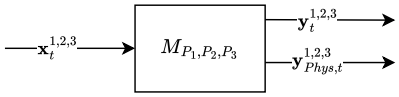


Figure 10. Schematic representation of PRNN for neural state-space model for our testbed.

the model learned just using the physics by eliminating the error resulting from the suboptimally estimated parameters from the physics-based model. Comparing the performance of the data-driven model and PRNN from Table 1 and Table 3, it could be observed that the PRNN model helps ensure the model follows the underlying physics and, hence, could help improve the performance of the data-driven model. The difference in the performance of nominal models for Pumps 1 and 2 between the data-driven model and PRNN is not significant due to the simplicity of the components. However, the improvement is evident in the case of Pump 3, which was set slightly differently to drop its performance instantaneously after 3 minutes and then behave normally again. In this case, the data-driven model's performance degrades as this noisy instance hurts the model's training. However, since the PRNN obtained the signal from the physics model, it was able to stay on track with the training process. Further, based on Table 3, and Figure 11, it is clear that the adaptation technique successfully adapts the learned PRNN model to faulty scenarios.

To estimate the amount of Carbon Dioxide (CO_2) produced by the cloud or personal computing resources used to execute the code, one may use the `Code Carbon`[2] library. It helps us track the carbon emissions for any computational process by considering the region where the machine is located, the amount of CPU and GPU consumed, the power used to run the particular process, and the overall machine's power consumption. By tracking all of this, the library can compute an estimate of the carbon intensity and energy consumption, thus resulting in the final number representing the CO_2 emissions.

Based on this library, Table 4 summarizes the power used as well as the CO_2 emissions in the process to train the model, retrain it using the standard transfer learning approach, and the adaptation of the model using the JFR method. For conducting this study, experiments were hosted to the Google Cloud Platform so as to have an accurate track of the compute resources as well as the carbon intensity. Using the local computer, the variables such as the particular power source

[2] `https://codecarbon.io/`

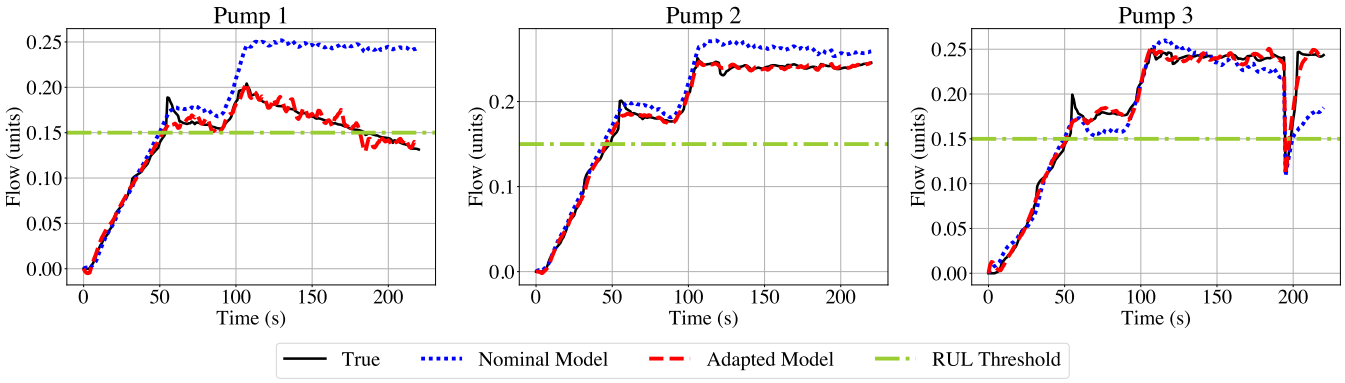


Figure 11. Flow estimates for the three pumps of the system representing the oilfield. The estimates are derived after running the forward Euler simulation on the PRNN model. The blue dotted line depicts the estimates derived using the model before adaptation (i.e.) in the nominal state. The red dashed line represents the estimates from the model after adaptation, and the black solid line represents the actual system behavior when the fault has been introduced in Pump 1. The green dashdot horizontal line represents the threshold that could be used to determine the RUL of the system.

being used and other local factors can skew the results.

Table 4. Summary of carbon emissions and power usage.

Phase	Carbon Emissions	Energy Consumed
Training	1.70×10^{-4}	2.84×10^{-3}
Retraining	1.70×10^{-4}	2.8×10^{-3}
Adaptation	1.37×10^{-6}	2.28×10^{-5}

From the results in Table 4, it is evident that the process of adaptation results in far lower levels of carbon emissions as well as less power usage. This is a clear advantage of using the adaptation technique instead of the standard retraining-based transfer learning as the carbon emissions are reduced, improving the sustainability aspects of the developed solution; the duration for which both processes are run is also much different where adaptation is approximately 10 times faster and uses far less computational resources.

4. Conclusions and Future Work

In this paper, we presented an automated online model adaptation framework for robust RUL prediction. We showcased the JFR-based adaptation technique to adapt the models representing the system for online RUL prediction and also extended this technique to adapt the hybrid ML approaches that provide robust system representation. The results indicate that this approach is much more computationally efficient than retraining a data-driven model based on standard transfer learning methods. In the future, we would like to continue modifying the algorithms to relax some of the assumptions. We would also like to extend this approach to switched hybrid systems that combine discrete modes along with continuous system dynamics in each discrete mode.

References

Cheng, H., Kong, X., Wang, Q., Ma, H., Yang, S., & Chen, G. (2023). Deep Transfer Learning Based on Dynamic Domain Adaptation for Remaining Useful Life Prediction Under Different Working Conditions. *Journal of Intelligent Manufacturing*, *34*(2), 587-613.

da Costa, P. R. d. O., Akçay, A., Zhang, Y., & Kaymak, U. (2020). Remaining Useful Lifetime Prediction via Deep Domain Adaptation. *Reliability Engineering & System Safety*, *195*, 106682.

Ding, Y., Ding, P., Zhao, X., Cao, Y., & Jia, M. (2022). Transfer Learning for Remaining Useful Life Prediction Across Operating Conditions Based on Multi-source Domain Adaptation. *IEEE/ASME Transactions on Mechatronics*, *27*(5), 4143-4152.

Forgione, M., Muni, A., Piga, D., & Gallieri, M. (2023). On the Adaptation of Recurrent Neural Networks for System Identification. *Automatica*, *155*, 111092.

Huang, Z., Xu, Z., Wang, W., & Sun, Y. (2015). Remaining Useful Life Prediction for a Nonlinear Heterogeneous Wiener Process Model With an Adaptive Drift. *IEEE Transactions on Reliability*, *64*(2), 687-700.

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017). Physics-Guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv preprint arXiv:1710.11431*, *2*.

Lei, Y., Li, N., Gontarz, S., Lin, J., Radkowski, S., & Dybala, J. (2016). A Model-Based Method for Remaining Useful Life Prediction of Machinery. *IEEE Transactions on Reliability*, *65*(3), 1314-1326.

Liu, J., Saxena, A., Goebel, K., Saha, B., & Wang, W. (2010). An Adaptive Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-ion Batteries. In *Annual conference of the PHM Society* (Vol. 2).

Liu, L., Guo, Q., Liu, D., & Peng, Y. (2019). Data-Driven

Remaining Useful Life Prediction Considering Sensor Anomaly Detection and Data Recovery. *IEEE Access*, *7*, 58336-58345.

Ma, M., & Mao, Z. (2021). Deep-Convolution-Based LSTM Network for Remaining Useful Life Prediction. *IEEE Transactions on Industrial Informatics*, *17*(3), 1658-1667.

Pan, D., Li, H., & Wang, S. (2022). Transfer Learning-Based Hybrid Remaining Useful Life Prediction for Lithium-Ion Batteries Under Different Stresses. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-10.

Sheth, P., Roychoudhury, I., Chatar, C., & Celaya, J. (2022). A Hybrid Physics-Based and Machine-Learning Approach for Stick/Slip Prediction. In *SPE/IADC Drilling Conference and Exhibition.*

Si, X.-S., Hu, C.-H., Chen, M.-Y., & Wang, W. (2011). An Adaptive and Nonlinear Drift-based Wiener Process for Remaining Useful Life Estimation. In *2011 Prognostics and System Health Management Conference* (p. 1-5).

Siahpour, S., Li, X., & Lee, J. (2022). A Novel Transfer Learning Approach in Remaining Useful Life Prediction for Incomplete Dataset. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-11.

Sun, C., Ma, M., Zhao, Z., Tian, S., Yan, R., & Chen, X. (2019). Deep Transfer Learning Based on Sparse Autoencoder for Remaining Useful Life Prediction of Tool in Manufacturing. *IEEE Transactions on Industrial Informatics*, *15*(4), 2416-2425.

Wang, Y., Zhao, Y., & Addepalli, S. (2020). Remaining Useful Life Prediction using Deep Learning Approaches: A Review. *Procedia Manufacturing*, *49*, 81-88.

Zhang, A., Wang, H., Li, S., Cui, Y., Liu, Z., Yang, G., & Hu, J. (2018). Transfer Learning with Deep Recurrent Neural Networks for Remaining Useful Life Estimation. *Applied Sciences*, *8*(12).

Zhang, Y., Xiong, R., He, H., & Liu, Z. (2017). A LSTM-RNN Method for the Lithuim-ion Battery Remaining Useful Life Prediction. In *2017 Prognostics and System Health Management Conference (PHM-Harbin)* (p. 1-4).

Zhang, Y., Yang, Y., Xiu, X., Li, H., & Liu, R. (2021). A Remaining Useful Life Prediction Method in the Early Stage of Stochastic Degradation Process. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *68*(6), 2027-2031.

BIOGRAPHIES

Prasham Sheth is a Data Scientist at SLB Software Technology Innovation Center. His research interests include the application of machine learning, deep learning, and hybrid modeling-based approaches to solving complex problems in computer vision and time-series analysis. He received a Master of Science in Data Science from Columbia University, New York, New York, USA, and a Bachelor of Technology in Computer Engineering from Nirma University, Ahmedabad, Gujarat, India.

Indranil Roychoudhury is a Principal AI Scientist at SLB Software Technology Innovation Center, and his primary area of research is time-series analysis by combining physics-based approaches with machine learning approaches. He holds his Ph.D. and M.S. in Computer Science from Vanderbilt University and was a Senior Research Scientist at NASA Ames Research Center before joining SLB. He is a Fellow of the Prognostics and Health Management Society and a Senior Member of IEEE.

Simulation-based remaining useful life prediction of rolling element bearings under varying operating conditions

Seyed Ali Hosseinli[1, 2], Ted Ooijevaar[3] and Konstantinos Gryllias[1, 2]

[1] *Department of Mechanical Engineering, KU Leuven*
[2] *Flanders Make @ KU Leuven*
Celestijnenlaan 300, BOX 2420, 3001 Leuven, Belgium
[3] *Flanders Make vzw, CoreLab MotionS, 3001 Leuven, Belgium*
konstantinos.gryllias@kuleuven.be

ABSTRACT

Remaining useful life (RUL) prediction of rolling element bearings is a complex task in the frame of condition monitoring which brings cost benefits to the industry by reducing unexpected downtimes and failures. Data-driven approaches based on deep learning have demonstrated exceptional performance in estimating RUL effectively. Nevertheless, challenges such as data scarcity for model training and varying operating conditions add more complexity to prognostic tasks using these methods. This study proposes a methodology for simulating the vibration signals during the degradation process of bearings in order to mitigate the need for historical data for training the models. Simulations are realized using a phenomenological model whose free parameters are adapted based on real measurements so that the simulated run-to-failure datasets are under the same influence of speed as the real dataset with almost the same degradation rate. The simulated dataset is used for model training. Moreover, the proposed methodology is able to react to the shaft speed and be flexible at the predictions when the speed of the bearing varies. The proposed model can take extra information regarding the operating speed and the sequential ordering of the measurements to be aware of the working conditions and the dynamics of the damage progression. The positive effect of the extra information is shown in the results. Model training is based on an unsupervised domain adaptation approach to reduce domain discrepancy between the simulated and real feature space. The effectiveness of the proposed method is examined according to bearing run-to-failure tests under varying operating conditions.

1. INTRODUCTION

Improving the accessibility of industrial assets is a crucial factor in boosting productivity and efficiency, leading to cost benefits for industries. This is achieved by exploiting the full life of components and avoiding premature replacements. Rolling element bearings, being the key component of rotary equipment in the industry, are prone to failures due to their frequent operation in harsh and demanding conditions, including high temperatures, heavy loads, and contaminated surroundings, which increase the possibility of unexpected failures. Failures could spread to the entire machine and lead to unplanned downtimes (Buzzoni et al., 2020; Tajiani & Vatn, 2023).

RUL estimation techniques can be employed to determine the remaining time until the failure occurs. Failure is defined as a state in which a health indicator crosses a predefined threshold such that the component is no longer able to operate in the desired way (Lei et al., 2018). Numerous methods can be employed to achieve this objective. However, data-driven approaches, particularly those based on deep learning, have recently demonstrated outstanding results due to their capability of modeling processes with high complexity like degradation process in bearings (Fink et al., 2020). The main bottleneck of using deep learning methods is the necessity of having a large amount of labeled pre-recorded run-to-failure data for training the models. This process is time-consuming, labor-intensive, and crucially, in industrial settings, failures may be infrequent, and maintenance is usually performed before failures (Arias Chao et al., 2022), which makes it hard to have a perfect dataset for training the models, since for model training, datasets should cover the whole life of bearings until the point of failure, and the degradation process is normally a long process that could take months or even years (Chen et al., 2023). Additionally, labeled data from

different bearings or machines operating under different conditions may be insufficient due to different deterioration trajectories and conditional probabilities. As a solution, leveraging simulated datasets for training deep learning models has surfaced as a practical approach to mitigate the constraints imposed by the limited availability of real labeled datasets. This approach enhances the overall performance of the models by providing a broader and more diverse set of training data (Hosseinli et al., 2023). Gryllias and Antoniadis (2012) generated artificial signals by a phenomenological model for different types of localized faults in bearings and then trained a Support Vector Machine (SVM) model using them. The real samples were then classified using the trained SVM model. Cui et al. (2020) proposed a method based on a 5-DOF dynamic model of bearings coupled with surface topography excitation to create a dictionary of many different degradation processes. Then, based on the similarity of the tested bearing and the simulated ones, the RUL of the tested bearing can be estimated based on the life label of the most matched sample. Deng et al. (2023) developed a 5-DOF dynamic model of bearings and generated a large amount of samples. Then, a particle filter-based dynamic calibration method was used to calibrate the parameters of the model based on observations. The simulated dataset was further used to train a deep learning model and estimate the RUL of real samples. Ai et al. (2023) utilized a phenomenological model to create a dataset for three types of fault: ball, inner race, and outer race for fault diagnosis. A deep learning model based on the transfer learning approach was then trained to remove the gap between the distributions of the real and simulated signals for fault classification of the real dataset.

Moreover, varying operating conditions, which can be seen in industrial cases such as wind turbines, servo motors, compressors, etc., pose another challenge for estimating the RUL of bearings. Developing a RUL prediction method that can respond to the operating conditions is of high importance since the developed models based on the assumption of steady operating conditions could not have satisfying performance under varying operating conditions (Chi et al., 2022; Liao & Tian, 2013). (Wang et al., 2021) developed a model-based method for RUL estimation by considering the joint dependency of degradation rate and time-varying operating conditions. The parameters of a system state function and an observation function were then estimated to model the degradation process of the system and predict the RUL of bearings. (Li et al., 2019) developed a state-space model for systems working under varying operating conditions. The model considered two effects of the varying operating conditions: changes in degradation rates and jumps in degradation signals. By estimating the underlying system state and the remaining time until it reaches a failure predefined threshold, the RUL of tested bearings was estimated. Zhang et al. (2022) proposed a normalization method that recalibrates the upward and downward abrupt jumps of sensor readings at the operational conditions change points. Then, the normalized sensor features and operating condition features were fed to a gated recurrent unit (GRU) to estimate the RUL of the aircraft turbofan engine dataset provided by NASA.

Motivated by the observation that the literature lacks a comprehensive exploration of RUL estimation under varying operating conditions using deep learning, in this paper, the proposed methodology consists of different steps including data simulation as a way of mitigating the influence of data scarcity and then a deep learning model based on a domain adaptation approach which gets the raw vibration signals as input as well as supplementary information on working conditions in which the signals are acquired in order to make the model aware of varying operating conditions. The rest of the paper can be summarized as follows:

1. Utilize a phenomenological model that simulates the general vibration signals of the bearings under different fault modes: ball, inner race, and outer race defects.
2. Adapt the phenomenological model based on the healthy real signals to tune the dynamic parameters of the model and also identify the effect of varying speed conditions on the amplitude of vibration signals.
3. Separate the effect of speed from the peak-to-peak health indicator so that it only indicates the degradation process which is void of the influence of speed and realizing anomaly detection based on this new health indicator (normalized peak-to-peak).
4. Realize curve-fitting on the normalized peak-to-peak after the anomaly to find out how fast the damage is progressing and then generate many synthetic run-to-failure data under the same influence of speed and damage progression as the real data to create a big training dataset for training a deep learning model.
5. Train a deep learning model according to the domain adversarial method to decrease the discrepancy between the unlabeled real data and the labeled simulated data. The deep learning model takes two additional inputs: speed information and sequence information in order to better understand the working conditions and the sequential relationship between each measurement.
6. Estimate the RUL of the real measurements using the trained deep-learning model.

In other words, the proposed methodology, as shown in Figure 1, is a digital twin (DT) that requires no historical data and is able to adapt itself to different rotating speeds, fault modes, and degradation rates of bearings.

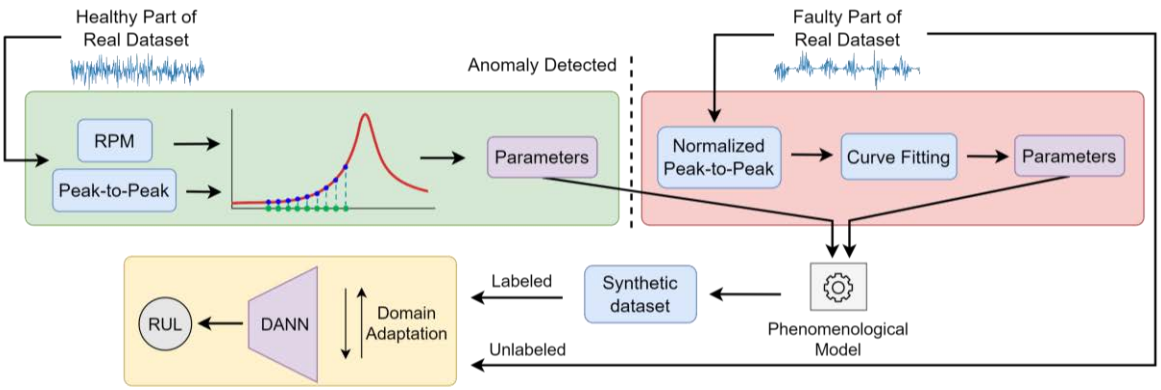


Figure 1. Pipeline of the proposed methodology

Peak-to-peak values are used here as a reference health indicator to tune the digital twin, since the EoL criterion is assumed to be defined on the peak-to-peak, and the synthetic dataset needs to be created under the same definition as the real dataset's EoL to be able to mimic historical datasets for training. Moreover, a few unlabeled real data that comes after the anomaly detection are used for unsupervised domain adaptation. They are unlabeled because their corresponding RULs are not known at this stage.

The rest of the paper is organized as follows. First, the fundamental theoretical background used in the proposed methodology is shortly introduced in Section 2. Moreover, the proposed approach to adapt the DT and predict the RUL is introduced in Section 3. Furthermore, a run-to-failure dataset captured under varying speed operating conditions is presented in Section 4, the methodology is applied, and the effectiveness of the proposed approach is demonstrated. Finally, Section 5 provides the conclusion of the paper.

2. THEORETICAL BACKGROUND

2.1. Phenomenological model

The phenomenological simulation of bearing vibration signals involves replicating the actual vibration signals of a real bearing. The simulation method achieves this by comparing the entire bearing and its supporting structure to a single-degree-of-freedom (SDOF) vibration system and introducing consecutive impulses to excite the structure, mirroring the effects of localized faults within the bearing. This approach allows for a representation that emulates the characteristics of real-world bearing vibration signals. The initial idea was proposed by McFadden and Smith (1984) and then it was improved by Antoni (2007) in order to have a more realistic spectral analysis. The simulated vibration signal can be generated by the following formula:

$$x_k(t) = S(k)\,.\,D(k)\sum_{i=-\infty}^{+\infty} h(t - iT - \tau_i)q(iT)A_i + n(t) \qquad (1)$$

where $S(k)$ and $D(k)$ are the amplitude modifiers regarding the speed and damage influence on the amplitude of the signals, respectively, for the k-th simulated signal in a degradation process. $h(t)$ is the impulse response of the equivalent SDOF system. T is the time between two consecutive impacts. i is the index of the i-th impact due to the fault, $n(t)$ accounts for the possible noise presented in the signals, and q is the amplitude modulating function due to the load distribution. A and τ are the parameters in order to take into account the randomness of the impact intensities and the moments that the impacts occur, respectively. According to (Antoni, 2007):

$$E\{\tau_i\tau_j\} = \delta_{ij}\sigma_\tau^2$$
$$E\{A_i^2\} = 1 + \delta_{ij}\sigma_A^2 \qquad (2)$$

where σ_τ and σ_A are the standard deviations, and δ_{ij} is the Kronecker symbol. The time period between two consecutive impacts depends on the rotational speed of the inner race of the bearing, and the mean value of the time interval ΔT is expressed by:

$$E\{\Delta T\} = \frac{E\{\Delta\theta\}}{2\pi f_r} \qquad (3)$$

where f_r is the inner race rotational speed and $\Delta\theta$ is the angular distance between two consecutive impacts which its mean value is expressed by:

$$E\{\Delta\theta\} = \frac{2\pi}{O_{imp}} \qquad (4)$$

where O_{imp} is the characteristic fault order, and it is defined as follows for different types of faults:

$$\begin{aligned}
&\text{Outer race} && \frac{n}{2}\left(1 - \frac{d}{D}\cos(\beta)\right)\\
&\text{Inner race} && \frac{n}{2}\left(1 + \frac{d}{D}\cos(\beta)\right)\\
&\text{Rolling element} && \frac{D}{2d}\left(1 - \left(\frac{d}{D}\cos(\beta)\right)^2\right)\\
&\text{Cage} && \frac{1}{2}\left(1 - \frac{d}{D}\cos(\beta)\right)
\end{aligned} \qquad (5)$$

where n is the number of rolling elements in the bearing, D is the pitch circle diameter, d is the bearing roller diameter and β represents the contact angle.

2.2. Domain adaptation

Acknowledging that simulated signals are derived from a simplistic model, incapable of capturing all aspects of faulty bearings or the degradation process, a distribution mismatch between real and simulated signals arises. This mismatch poses challenges in generalization when deploying trained models on real datasets. To tackle this issue, a domain adaptation method is employed to enhance generalization by transferring knowledge acquired from the source domain $\mathcal{D}_S$, where simulated signals originate, to the target domain $\mathcal{D}_T$, representing real-world datasets (Pan & Yang, 2010). This facilitates improved performance and adaptability of trained models. In this case, the marginal probability distributions of source and target domains, $P(x_S)$ and $P(x_T)$, are assumed to be different due to the simplicity of the simulations, but their conditional probability distributions, $P(y_S|x_S)$ and $P(y_T|x_T)$, are assumed to be the same due to the fact that in the preprocessing steps, the digital twin is tuned based on the real data in terms of the type of fault, the influence of speed, and the dynamic characteristics of the bearing.

This paper employs a Domain Adversarial Neural Network (DANN) to tackle the abovementioned domain shift problem. The network, illustrated in Figure 1, comprises three key components: a feature extractor G_f a domain classifier G_d, and a label predictor or regressor G_r. The feature extractor G_f is typically a deep neural network responsible for learning high-level representations from input data. It transforms input samples into a latent representation encoding valuable features for subsequent layers. The domain classifier G_d is another neural network component that predicts the domain of input samples based on extracted features, aiming to differentiate between the source and target domains. During training, the domain classifier seeks to maximize its accuracy, while the feature extractor aims to minimize this accuracy by gradient reversal. This adversarial training process results in the domain classifier being unable to distinguish features from different domains, indicating that the feature extractor can extract domain-invariant features. Additionally, the label predictor or regressor layer G_r utilizes these domain-invariant features to estimate the output, contributing to the overall goal of addressing the domain shift problem (Ganin et al., 2016). The objective function of the model is:

$$\mathcal{L}(\theta_f,\theta_r,\theta_d) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_r^i(\theta_f,\theta_r) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d^i(\theta_f,\theta_d) + \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d^i(\theta_f,\theta_d)\right) \quad (6)$$

where n and n' are the number of samples presented in the source domain and the target domain datasets respectively, and λ is a hyperparameter that controls the trade-off between the regression loss and the domain adversarial loss during training. $\mathcal{L}_r$ and $\mathcal{L}_d$ are defined as:

$$\mathcal{L}_r^i(\theta_f,\theta_r) = \mathcal{L}_r\big(G_r\big(G_f(x_i;\theta_f);\theta_r\big),y_i\big)$$
$$\mathcal{L}_d^i(\theta_f,\theta_d) = \mathcal{L}_d\big(G_d\big(G_f(x_i;\theta_f);\theta_d\big),d_i\big) \quad (7)$$

where θ_f, θ_r, and θ_d are the trainable parameters of the G_f, G_r, and G_d respectively.

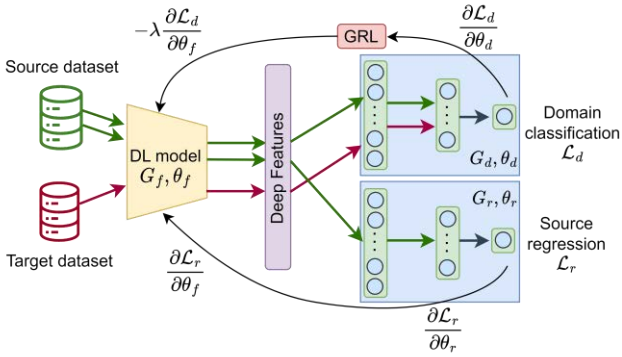


Figure 2. DANN architecture for regression task

3. Proposed Approach

To create a synthetic run-to-failure dataset by the digital twin, at first the dynamic characteristics of the phenomenological model should be tuned based on the real data. Then, the modifier functions $S(k)$ and $D(k)$, introduced in Section 2, are determined.

3.1. Dynamic characteristics

Given the fact that rolling element bearings vibrate even in healthy conditions due to the waviness of the surfaces and other sources of imperfections (Harsha et al., 2003; Jawad & Jaber, 2022), the resonance frequency of the structure can still be seen in the frequency content of the vibration signals (Ghafari et al., 2010). Therefore, by considering the Fast Fourier Transform (FFT) of the healthy signals, the dominant natural frequency of the structure, ω_n, can be found and then used in the phenomenological model as an equivalent SDOF system. Moreover, the logarithmic decrement can be used to see at which rate the amplitude of the impact responses, x, in real measurements is decreasing in order to determine ζ.

$$h(t) = e^{-\zeta\omega_n t}\sin\left(\sqrt{1-\zeta^2}\omega_n t\right)$$
$$\zeta = \frac{\delta}{\sqrt{4\pi^2+\delta^2}} \quad (8)$$
$$\delta = ln\left|\frac{x_1}{x_2}\right|$$

3.2. Influence of speed, $S(k)$

The real signals before the detection of the anomaly can be used to recognize the influence of speed on the amplitude of the vibration signals. Figure 3 (a) shows a speed profile and its effect on the peak-to-peak amplitude of the vibration. The important point here is to consider the possibility of the structure resonance when speed is varying. In other words, increasing speed does not necessarily result in an increasing amplitude. Figures 3 (b) and (c) show two possible behaviors that can be seen in speed-varying scenarios. Increasing amplitude with increasing speed can be a sign of crossing no resonance frequency in that specific speed range (Salunkhe & Desavale, 2021).

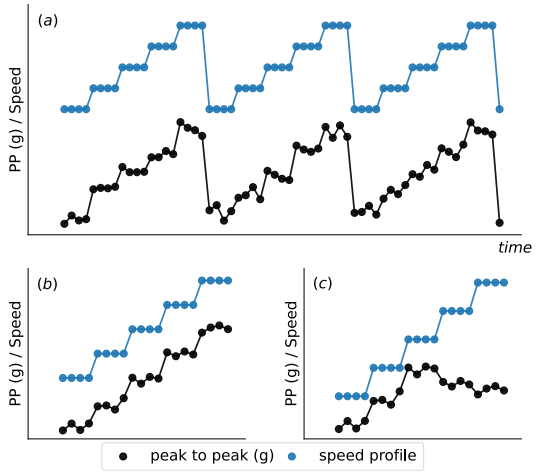


Figure 3. (a) Varying speed effect on the peak-to-peak amplitude of the signals, (b) in case of no resonance crossing, (c) in case of crossing a resonance frequency

Corresponding to each operating speed, a constant parameter c can be found to create a link between the rotating speed and the vibration amplitude or the peak-to-peak:

$$P_j = c_j \,.\, rpm_j \tag{9}$$

where P is the peak-to-peak of real signals, c is a constant parameter, rpm is the operating speed, and j is the index of measurements. In this way, any non-linearity between speed and vibration due to the frequency response of the structure can be captured (Figure 4).

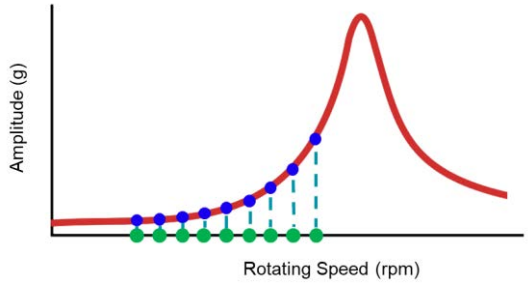


Figure 4. Capturing the complex relationship between speed and vibration amplitude in real measurements

It should be noted that these analyses must be done based on a few measurements at the beginning of the operation of the machine to be far from the influence of bearing faults. After the detection of the anomaly, when the synthetic run-to-failure dataset should be created, a speed profile should also be provided so that the digital twin can generate signals accordingly. Since a random speed profile might be desired at this stage, an interpolation would be needed to find the correct value of c while dealing with unseen speed values. Then, the modifier function $S(k)$ in equation 1 will be:

$$S(k) = c_k \,.\, rpm_k \tag{10}$$

3.3. Normalized health indicator

By knowing the relation between speed and vibration, the effect of speed can be removed from the peak-to-peak amplitude of the vibration signals by equation 11, meaning that any changes in the health indicator that are not associated with speed can manifest itself more clearly. This method will be used to find anomalies. The Normalized peak-to-peak amplitude, P_N, is constructed as follows:

$$P_{N,j} = \frac{P_j}{c_j \,.\, rpm_j} \tag{11}$$

Obviously, for the healthy samples before the detection of the anomaly, $P_N \sim 1$. Figure 5 (b) shows the normalized peak-to-peak amplitude whose mean and standard deviation in the time interval $[t_0, t_1]$, which is at the beginning of the measurements, can be used as the threshold for anomaly detection. $K_{anomaly}$ is used to refer to the k-th signal in the degradation process where the anomaly occurs. As shown in Figure 5 (b), the fluctuations caused by the varying speed profile no longer exist in the normalized peak-to-peak amplitude.

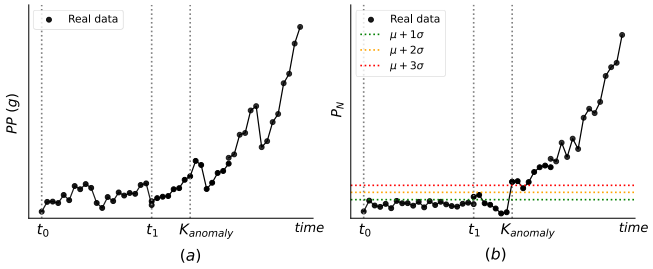


Figure 5. (a) Peak-to-peak amplitude of real data, (b) anomaly detection by the normalized peak-to-peak

3.4. Influence of damage, $D(k)$

After the anomaly, a limited number of signals will be used for curve fitting in order to estimate the degradation rate of the real bearing so that the digital twin will be able to generate a synthetic run-to-failure dataset with almost the same degradation rate as the real bearing. As shown in Figure 6 (a), curve fitting is done based on the normalized peak-to-peak

amplitude because the effect of speed has been removed, and it is only the degradation process that plays a role. The modifier function $D(k)$ in equation 1 can be modeled by an exponential function to approximate the degradation trajectory of the normalized peak-to-peak amplitude:

$$D(k) = e^{a(k-K_{anomaly})} \tag{12}$$

where a is a constant parameter that defines the degradation rate. By introducing slight variations in the parameter a in function $D(k)$, various degradation trajectories can be built in order to have a big synthetic run-to-failure dataset. The variations are such that the time difference between the synthetic EoLs is limited to the *Simulation range* as shown in Figure 6 (b). This figure shows the typical degradation trajectories of the peak-to-peak amplitude of the synthetic dataset generated by equation 1. Domain adaptation is also done using a few real unlabeled available measurements after the detection of the anomaly.

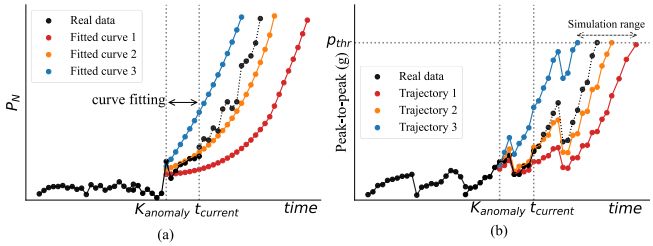


Figure 6. (a) Curve fitting according to the normalized peak-to-peak amplitude, (b) peak-to-peak amplitude of the synthetic dataset

3.5. Encoding

In order to encode the speed and sequence label of each measurement, which will be fed into the deep learning model, this paper adopted one of the well-known methods of information encoding from the natural language processing (NLP) research domain. Positional encodings are used to make the transformers aware of the relative or absolute order of the words inside a sentence (Vaswani et al., 2017). To encode the positional information, *sine* and *cosine* functions with different frequencies can be used:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{13}$$

where pos is the position of the word, d_{model} is the dimension of the word embeddings, and i represents the dimension of the positional encoding. Unique encoding for each position is achieved by using *sine* and *cosine* functions with varying frequencies, making the model able to distinguish the sequential order of measurements. It is important to highlight that each positional encoding with offset k, PE_{pos+k} can be described as a linear function of the positional encodings PE_{pos}, this characteristic enables the model to readily learn the relative dependencies between different positions, contributing to the model's ability to capture sequential information and relationships effectively (Vaswani et al., 2017).

The concept of positional encoding can be transferred to the prognosis research domain. The sequential order of vibration signals obtained from a bearing holds significant importance in prognosis, serving as an indicator of how the damage progresses over time. Moreover, equation 13 can be utilized for encoding speed information. While this encoding method might lack a direct physical interpretation, it serves the purpose of making the neural network aware of distinctions among vibration signals working in different conditions. Each operating condition should have a unique encoding by which the raw vibration signals are accompanied while feeding to the model. d_{model} is a hyperparameter that will be determined in section 4.1, and the value of pos is an integer number that starts from 1, representing the sequential order of each measurement. The same way is followed to encode the speed information. For example, $pos = 1$ is used for the lowest rotational speed. For each pos, the value of i starts from 0 and ends in $\frac{d_{model}}{2}$, forming a vector of length d_{model}. For each i there are two values, one from $sine$ function and the other from $cosine$ function. For example, the encoding for $pos = 1$ is $[PE_{(1,0)}, PE_{(1,1)}, \dots, PE_{(1,d_{model}-1)}]$ which is a one-dimensional vector.

3.6. RUL curve for varying speed scenario

One of the most important outcomes of the proposed methodology is to see how the speed is influencing the RUL. Obviously, for higher speeds, lower RUL is expected, and vice versa. To the best knowledge of the authors, no paper has taken into account the effect of speed on the RUL curve.

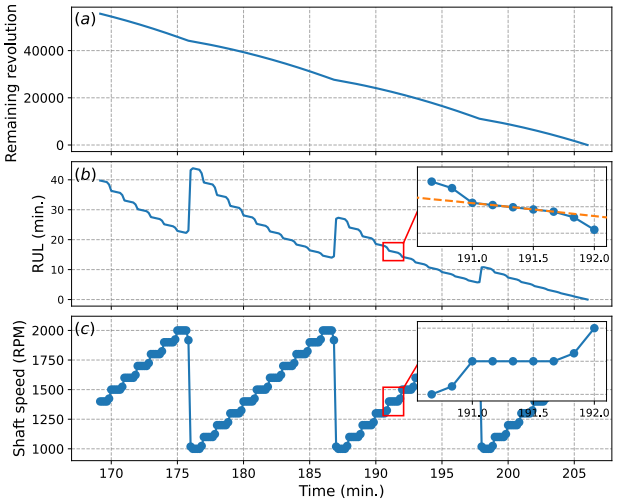


Figure 7. (a), (b) Remaining revolutions and remaining time after the detection of anomaly, (c) the corresponding speed profile

The proposed idea is to estimate the remaining revolutions until the end-of-life, and in the post-processing step, the number of revolutions can be transferred to the remaining time by simply dividing it by the operating speed. Figure 7 shows the remaining revolutions and the RUL for a measurement campaign which has been done under varying speed operating conditions. Equation 14 is used to calculate the total number of revolutions after the detection of anomaly which are then used as the labels in the model training.

$$Rev_{Total} = \sum_{i=K_{anomaly}}^{N} rpm_i . \Delta t$$

$$Rev_n = Rev_{Total} - \sum_{i=1}^{n} rpm_{i+K_{anomaly}} . \Delta t \qquad (14)$$

$$RUL_n = \frac{Rev_n}{rpm_{n+K_{anomaly}}}$$

where Rev_{Total} is the total number of revolutions after the detection of anomaly, N is the number of samples in the run-to-failure experiment, Δt is the length of each measurement. Rev_n and RUL_n are the remaining revolutions and the remaining time until the end-of-life for the n-th sample, respectively. Rev_n is used as the labels for model training.

The important point is that the slope of the RUL curve is -1 as long as the speed is constant, as shown in Figure 7 (b) by the orange line, preserving the most useful property of the RUL curve.

3.7. Deep-learning model

As mentioned in Section 2.2, a deep learning model based on the DANN model is used to estimate the RUL of the real bearings. Two supplementary information, speed and sequential ordering of the measurements, have been encoded and will be fed to the model as extra inputs in addition to the raw vibration signal. As shown in Figure 8, Convolutional Neural Network (CNN) is used to extract the local information and deep features automatically from the raw vibration signals. The extracted features are concatenated by two encoded inputs to form a bigger 1-D vector which is followed by two parallel Fully Connected (FC) layers, a domain discriminator, and a source regressor. The loss function of the regressor part is the mean squared error and the loss function of the discriminator part is the binary cross entropy which is expressed as follows:

$$\mathcal{L}_d = -y . log(\bar{y}) - (1 - y) . log(1 - \bar{y}) \qquad (15)$$

where $y \in \{0, 1\}$ is the domain label and $\bar{y}$ is the predicted value between 0 and 1. Table 1 and Table 2 show the network parameters in detail. As depicted in Figure 8, the gradient reversal layer (GRL) with the trade-off parameter $\lambda = 0.1$ is also added as the first layer of the discriminator part to reverse the gradient in the backpropagation process.

Table 1. Network parameters of the feature extractor

Layer	Type	Filter/ Kernel/Stride	Activation function
1	1D CNN	4/128/16	ReLU
2	Max Pooling	-/8/8	-
3	1D CNN	16/16/8	ReLU
4	Max Pooling	-/8/8	-

Table 2. FC parameters in the regressor and the discriminator

Regressor part		Discriminator part	
Layer	Units/Activation function	Layer	Units/Activation function
1	64/ ReLU	1	64/ ReLU
2	32/ ReLU	2	32/ ReLU
3	1/ ReLU	3	1/sigmoid

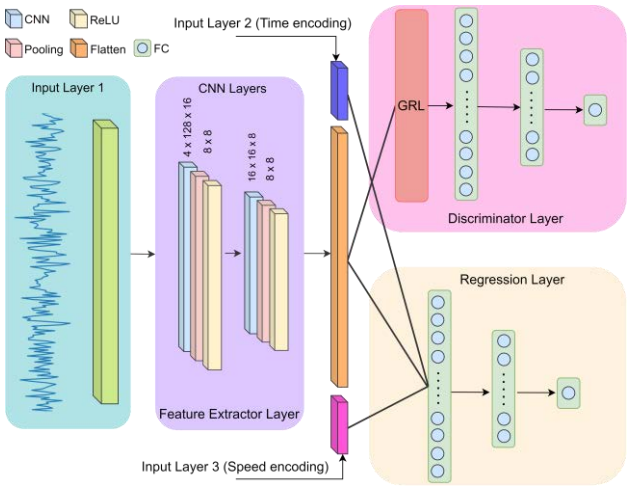


Figure 8. The architecture of the proposed model

It is important to emphasize that the length of the input signal must be sufficiently long to encompass an adequate number of impacts resulting from faults in the bearing. Notably, the number of impacts due to a ball defect in one revolution of the shaft is lower compared to other types of faults. Using equation 16, the number of data points needed to cover at least 1 impact due to the ball defect can be calculated. This number will cover more than one impact if a different type of defect is present at any speed (Hosseinli et al., 2023).

$$L_c = \frac{F_S}{BSF} \qquad (16)$$

where L_c is the critical length of the signal, BSF is the ball spin frequency at the lowest shaft speed, and F_S is the sampling frequency.

4. Application of the Methodology and Results

4.1. Case study

Smart Maintenance (SM) dataset provided by Flanders Make (Ooijevaar et al., 2019) consists of accelerated life tests where indentations were deliberately created on the inner races (IR) of bearings using a Rockwell-C indenter with a force of 100 kg before the tests started to run. The radial load is 9 kN and the rotational speed follows a periodic stepwise profile starting from 1000 rpm to 2000 rpm, each step is 100 rpm and is maintained for 60 seconds. The type of test bearings is 6205-C-TVH from FAG. The sampling rate frequency is 50 kHz, and a peak-to-peak amplitude of 15g is considered the end-of-life criterion in this study. Figure 9 shows the peak-to-peak amplitude and the speed profile of one of the measurement campaigns. Table 3 shows the specifications of all the bearings used in this study.

Table 3. Bearing information in the SM dataset

Bearing	Test duration	Anomaly detected at	Fault
A148	142.5 min.	112.6 min.	IR
A150	197.5 min.	169.1 min.	IR
A153	229.3 min.	207.8 min.	IR
A154	126.0 min.	98.8 min.	IR
A155	369.3 min.	348.6 min.	IR
A156	251.3 min.	224.0 min.	IR

Referring to equation 16, a signal of 25000 data points is set as the input to make sure that at least 20 impacts will be covered in the critical scenario for the Smart Maintenance dataset. The length of the encodings, d_{model} in equation 13, should be kept lower than the length of the deep features after the Flatten layer. This ensures that subsequent layers can effectively learn the deep features by maintaining a lower dimensionality for these encodings compared to the deep features, the model can efficiently process and integrate additional information without overwhelming the learning process or introducing unnecessary complexity. For the architecture described in Table 1, the length of the deep features for the input length of 25000 is 112.

Table 4. Length of the inputs of the proposed architecture

Input No.	Length
Input 1 (raw signal)	25000
Input 2 (time encoding)	24
Input 3 (speed encoding)	24

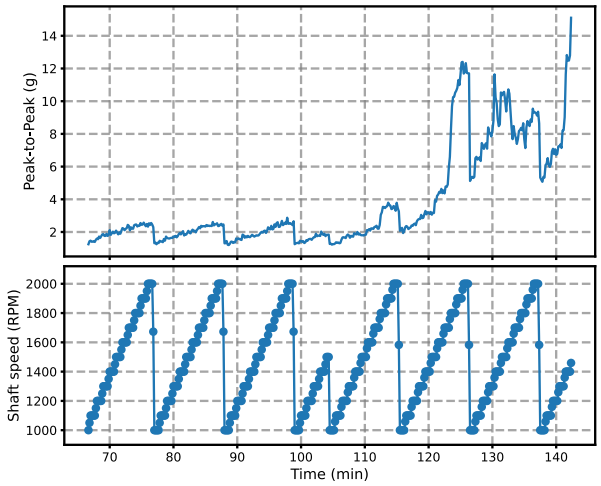


Figure 9. Bearing A148, peak-to-peak amplitude and speed profile

4.2. Results and discussions

Before the anomaly occurs, the relation between speed and vibration amplitude is analyzed. After the anomaly, curve fitting is done based on the available data, as shown in Figure 12. 10 minutes of measurements are used at this stage. This few unlabeled available real data is also used for domain adaptation while model training. It should be highlighted that the unlabeled past samples will be labeled in the inference stage. Despite the passage of time, labeling the past samples is still valuable, since it indicates what were the predictions from a few moments ago which can be used for decision-making. Figure 10 shows the result of anomaly detection and the corresponding speed profile for bearing A148.

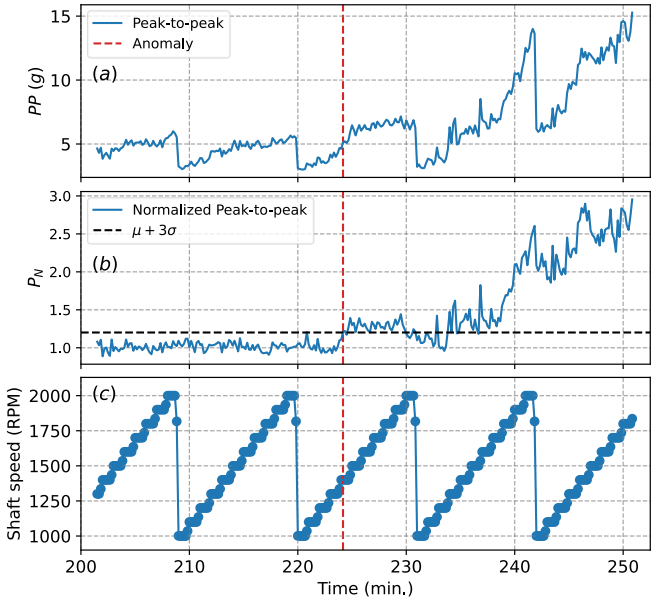


Figure 10. Bearing A156, (a) Peak-to-peak amplitude, (b) Normalized peak-to-peak amplitude, (c) Corresponding speed profile

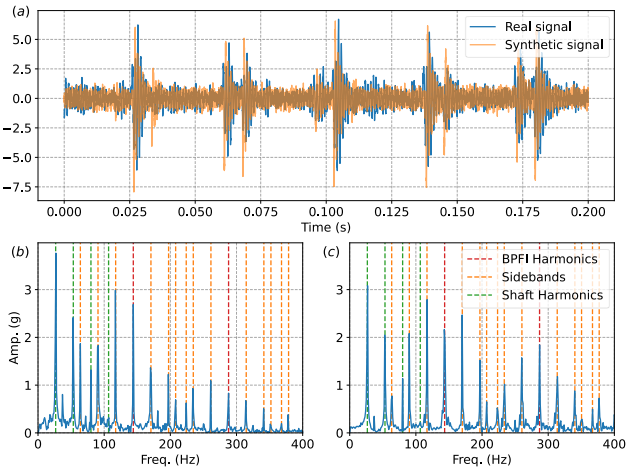


Figure 11. Bearing A148; (a) Comparison between the real and generated signals, (b) envelope spectrum of the real signal, (c) generated signal

Figure 11 shows one of the examples of the generated signals for bearing A148 after the anomaly. The results of the proposed method for anomaly detection are mentioned in Table 3.

After adapting the digital twin in terms of the speed influence and degradation rate, 6 trajectories are generated as shown in Figure 12. Their corresponding raw vibration signals will be the input of the deep learning model for training. The *Simulation range* is chosen to be 40 minutes. The synthetic run-to-failure dataset is used to train 5 models. A simple CNN model that neither includes the discriminator part of the architecture nor encodings, a DANN model without encodings, proposed model 1 with CNN and only speed encodings, proposed model 2 with CNN and only sequence encodings, and proposed model 3 with DANN and both speed and sequence encodings. Table 5 shows the superior performance of the proposed model 3 which in all cases can improve the root mean squared error, RMSE, of the RUL predictions compared to the DANN and CNN model. As discussed before, the important point of feeding the operating condition and the sequential information to the models is to make the model aware of the working environment and any other information that influences the physical behavior of the assets. This fact is perfectly shown in Figure 13 where by using the t-distributed stochastic neighbor embedding (t-SNE) technique the feature distribution of the second to the last fully connected layer in the regressor part of the proposed model is visualized. This layer outputs a 32-dimensional feature space that t-SNE can reduce the dimension to a lower one such as a 2-dimensional feature space which is easier to visualize. Figure 13 shows how the extra information fed to the model helps to distinguish between different speeds, resulting in better predictions. More importantly, supplementary information fed to the model makes the model more robust against the major changes in the speed profile. For example, bearing A156 underwent two major changes in operating speed after the detection of the anomaly. As depicted in Figure 14, abrupt speed changes from 2000 rpm to 1000 rpm in a short time interval led the CNN and DANN models to have a higher error in the predictions. Proving that these models have less control over the predictions when speed plays an impactful role. The proposed model makes satisfying predictions at the moment of abrupt speed changes and the predicted RUL is not too far from the ground truth, showing that the proposed model understands the relationship between rotating speed, vibration, and degradation severity thanks to the encodings. Most importantly, the estimated RUL is reactive to the operating speed. Higher speeds lead to lower RUL and vice versa. This property of the proposed method makes it applicable to real industrial cases where a varying speed profile is used.

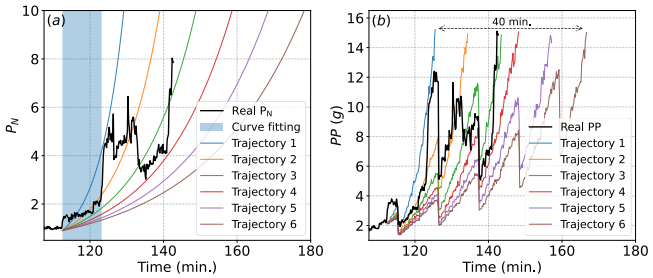


Figure 12. Bearing A148; (a) several trajectories by curve fitting on the real normalized peak-to-peak, (b) peak-to-peak of the generated signals by digital twin

Table 5. RMSE of the predicted RUL of the SM bearings in minutes

Bearing	CNN	DA NN	Proposed model 1	Proposed model 2	Proposed model 3
A148	6.6	7.8	7.3	**5.7**	5.8
A150	7.6	7.9	6.5	**5.7**	**5.7**
A153	4.2	3.8	2.8	2.5	**2.4**
A154	6.2	6.0	5.2	**2.7**	**2.7**
A155	5.5	5.3	4.7	3.7	**2.9**
A156	7.4	7.0	4.5	3.6	**2.7**

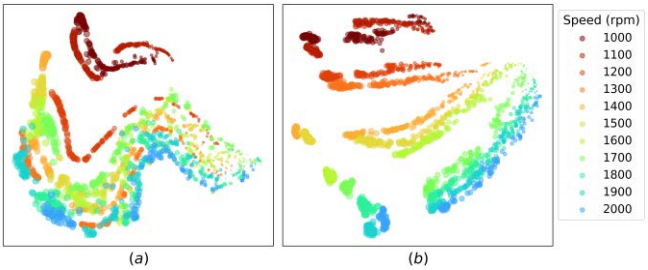


Figure 13. t-SNE visualization of the second to the last layer of the regressor part, the size of circles is proportional to the RUL, bearing A156, (a) CNN, (b) Proposed model 3

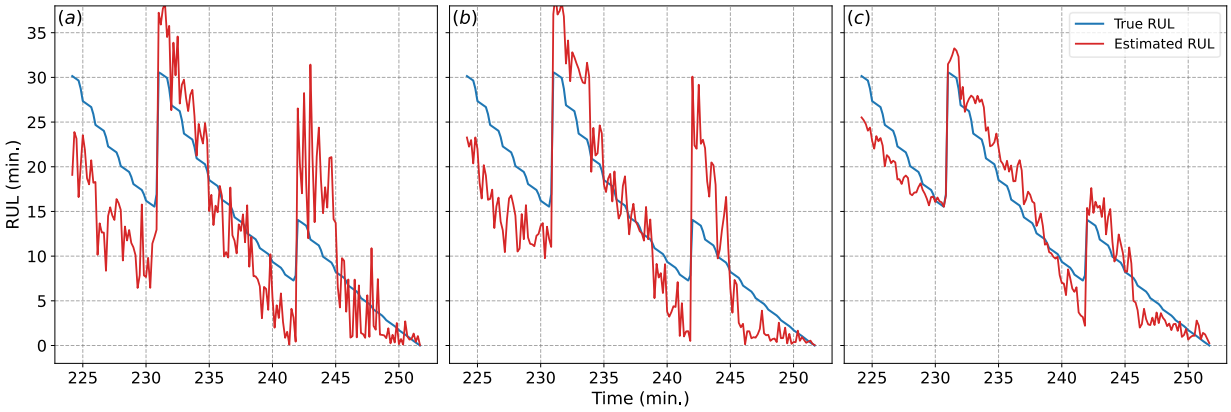


Figure 14. Predicted RUL of the bearing A156, (a) CNN, (b) DANN, (c) Proposed model 3

5. Conclusion

The objective of this study is to create a digital twin based on the simple physical knowledge of the fault progression phenomena. Utilizing a phenomenological model to generate synthetic signals helps to have a big synthetic run-to-failure dataset under varying speed operating conditions for training the machine learning models. On the other hand, leveraging the phenomenological model and simulated signals enhances the cost-effectiveness of the proposed approach by minimizing the reliance on historical run-to-failure datasets. The proposed methodology shows how the synthetic dataset should be adapted while facing varying speed scenarios. Additionally, the proposed model facilitates the integration of supplementary information regarding the working conditions and sequential ordering of measurements in a deep-learning model for prognosis and demonstrates that using extra information in the architecture of the DANN model enables the model to gain knowledge about both the operating conditions and the dynamics of damage progression. Moreover, a few unlabeled measurements from the real dataset after anomaly are used for domain adaptation in an adversarial way to reduce the gap between the feature distribution of the real and simulated dataset. Encoding the extra information, despite the lack of physical meaning, can aid the network in distinguishing signals from different operating conditions and identifying their relative relationships. Experimental results on the SM dataset demonstrate that the proposed model achieves improved RUL estimation accuracy, particularly in scenarios involving abrupt speed changes, and delivers more reliable predictions. The estimated RUL can also react to the operating speed which is a must in prognosis and decision making. Thanks to the t-SNE technique, the model's ability to discriminate between different operating conditions has been validated. The flexibility of the proposed method in recognizing the speed influences on the amplitude of the signals makes it applicable to the various speed profiles including random profiles, and also different speed ranges, whether or not they cross the resonance frequency of the structure. Experimental results have shown the capability of the proposed method compared to the models that do not utilize encodings.

Acknowledgment

The authors would like to acknowledge the support of Flanders Make, the strategic research center for the manufacturing industry in the context of the DGTwin Prediction project, the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program, and the Flemish Supercomputer Center (VSC) for the computation resource.

References

Ai, T., Liu, Z., Zhang, J., Liu, H., Jin, Y., & Zuo, M. (2023). Fully Simulated-Data-Driven Transfer-Learning Method for Rolling-Bearing-Fault Diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *72*. https://doi.org/10.1109/TIM.2023.3301901

Antoni, J. (2007). Cyclic spectral analysis of rolling-element bearing signals: Facts and fictions. *Journal of Sound and Vibration*, *304*(3–5), 497–529. https://doi.org/10.1016/j.jsv.2007.02.029

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering and System Safety*, *217*. https://doi.org/10.1016/j.ress.2021.107961

Buzzoni, M., D’Elia, G., & Cocconcelli, M. (2020). A tool for validating and benchmarking signal processing techniques applied to machine diagnosis. *Mechanical Systems and Signal Processing*, *139*, 106618. https://doi.org/10.1016/j.ymssp.2020.106618

Chen, J., Huang, R., Chen, Z., Mao, W., & Li, W. (2023). Transfer learning algorithms for bearing remaining useful life prediction: A comprehensive review from an industrial application perspective. *Mechanical Systems and Signal Processing*, *193*, 110239. https://doi.org/10.1016/j.ymssp.2023.110239

Chi, F., Yang, X., Shao, S., & Zhang, Q. (2022). Bearing Fault Diagnosis for Time-Varying System Using

Vibration–Speed Fusion Network Based on Self-Attention and Sparse Feature Extraction. *Machines*, *10*(10), 948. https://doi.org/10.3390/machines10100948

Cui, L., Wang, X., Wang, H., & Jiang, H. (2020). Remaining useful life prediction of rolling element bearings based on simulated performance degradation dictionary. *Mechanism and Machine Theory*, *153*, 103967. https://doi.org/10.1016/J.MECHMACHTHEORY.2020.103967

Deng, Y., Du, S., Wang, D., Shao, Y., & Huang, D. (2023). A Calibration-Based Hybrid Transfer Learning Framework for RUL Prediction of Rolling Bearing Across Different Machines. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–15. https://doi.org/10.1109/TIM.2023.3260283

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W. J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*. https://doi.org/10.1016/j.engappai.2020.103678

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). *Domain-Adversarial Training of Neural Networks*.

Ghafari, S. H., Abdel-Rahman, E. M., Golnaraghi, F., & Ismail, F. (2010). Vibrations of balanced fault-free ball bearings. *Journal of Sound and Vibration*, *329*(9), 1332–1347. https://doi.org/10.1016/j.jsv.2009.11.003

Gryllias, K. C., & Antoniadis, I. A. (2012). A Support Vector Machine approach based on physical model training for rolling element bearing fault detection in industrial environments. *Engineering Applications of Artificial Intelligence*, *25*(2), 326–344. https://doi.org/10.1016/j.engappai.2011.09.010

Harsha, S. P., Sandeep, K., & Prakash, R. (2003). The effect of speed of balanced rotor on nonlinear vibrations associated with ball bearings. *International Journal of Mechanical Sciences*, *45*(4), 725–740. https://doi.org/10.1016/S0020-7403(03)00064-X

Hosseinli, S. A., Ooijevaar, T., & Gryllias, K. (2023). Context-aware machine learning for estimating the remaining useful life of bearings under varying speed operating conditions. *Annual Conference of the PHM Society*, *15*(1). https://doi.org/10.36001/phmconf.2023.v15i1.3571

Jawad, S., & Jaber, A. (2022). Bearings Health Monitoring Based on Frequency-Domain Vibration Signals Analysis. *Engineering and Technology Journal*, *41*(1), 86–95. https://doi.org/10.30684/etj.2022.131581.1043

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. https://doi.org/10.1016/j.ymssp.2017.11.016

Li, N., Gebraeel, N., Lei, Y., Bian, L., & Si, X. (2019). Remaining useful life prediction of machinery under time-varying operating conditions based on a two-factor state-space model. *Reliability Engineering and System Safety*, *186*, 88–100. https://doi.org/10.1016/j.ress.2019.02.017

Liao, H., & Tian, Z. (2013). A framework for predicting the remaining useful life of a single unit under time-varying operating conditions. *IIE Transactions (Institute of Industrial Engineers)*, *45*(9), 964–980. https://doi.org/10.1080/0740817X.2012.705451

McFadden, P. D., & Smith, J. D. (1984). Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of Sound and Vibration*, *96*(1), 69–82. https://doi.org/10.1016/0022-460X(84)90595-9

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Salunkhe, V. G., & Desavale, R. G. (2021). An Intelligent Prediction for Detecting Bearing Vibration Characteristics Using a Machine Learning Model. *Journal of Nondestructive Evaluation, Diagnostics and Prognostics of Engineering Systems*, *4*(3). https://doi.org/10.1115/1.4049938

Tajiani, B., & Vatn, J. (2023). Adaptive remaining useful life prediction framework with stochastic failure threshold for experimental bearings with different lifetimes under contaminated condition. *International Journal of System Assurance Engineering and Management*, *14*(5), 1756–1777. https://doi.org/10.1007/s13198-023-01979-0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*.

Wang, H., Liao, H., & Ma, X. (2021). Remaining Useful Life Prediction Considering Joint Dependency of Degradation Rate and Variation on Time-Varying Operating Conditions. *IEEE Transactions on Reliability*, *70*(2), 761–774. https://doi.org/10.1109/TR.2020.3002262

Zhang, Z., Chen, X., & Zio, E. (2022). A framework for predicting the remaining useful life of machinery working under time-varying operational conditions[Formula presented]. *Applied Soft Computing*, *126*. https://doi.org/10.1016/j.asoc.2022.109164

BIOGRAPHIES

Seyed Ali Hosseinli received his BSc and MSc degrees in Railway Engineering from the Iranian University of Science and Technology (IUST) and in Mechanical Engineering from the Sharif University of Technology (SUT), respectively. He is currently a Ph.D. researcher at the Department of Mechanical Engineering of KU Leuven, Belgium. His research interests focus on machine learning techniques in prognostics and fault diagnosis of rotary equipment.

Ted Ooijevaar is a Monitoring Technology Domain Lead and Senior Research Engineer at Flanders Make since 2015. In this role, he defines, leads, and performs industry-driven applied research and development to support companies in the manufacturing industry. Ted received a Ph.D. degree in the field of Mechanical Engineering from the University of Twente and has special expertise in the field of sensing and monitoring, signal processing and data analytics, dynamics and mechanics, modeling, and experimental testing.

Konstantinos Gryllias holds a 5 years engineering diploma degree and a Ph.D. degree in Mechanical Engineering from the National Technical University of Athens, Greece. He holds an associate professor position on vibro-acoustics of machines and transportation systems at the Department of Mechanical Engineering of KU Leuven, Belgium. He is also the manager of the University Core Lab Flanders Make@KU Leuven - MPro of Flanders Make, Belgium. His research interests lie in the fields of condition monitoring, signal processing, prognostics, and health management of mech. & mechatronic system.

Soft Ordering 1-D CNN to Estimate the Capacity Factor of Windfarms for Identifying the Age-Related Performance Degradation

Manuel S Mathew[1], Surya Teja Kandukuri[2], Christian W Omlin[3]

[1,3]*University of Agder, Jon Lilletuns vei 9, 4879 Grimstad, Norway*
manuel.s.mathew@uia.no
christian.omlin@uia.no

[2]*Norwegian Research Centre, Energy & Technology Department, Tullins Gate 2, 0166 Oslo, Norway*
suka@norceresearch.no

ABSTRACT

Wind energy plays a vital role in meeting the sustainable development goals set forth by the United Nations. Performance of wind energy farms degrades gradually with aging. For deriving maximum benefits from these capital-intensive projects, these degradation patten should be analyzed and understood. Variations in the capacity factor over the years could be an indication of the age-related degradation of the wind farms. In this study, we propose a novel data-driven model to estimate the capacity factor of wind farms, which could then be used to estimate its age-related performance decline. For this, a 1-dimensional convolutional neural network (1-D CNN) is developed with a soft ordering mechanism under this study. The model was optimized using Huber loss to counteract the effects of outliers in data. The developed model could perform very well in capturing the underlying dynamics in the data as evidenced by a normalized root mean squared error (NRMSE) of 0.102 and a mean absolute error (MAE) of 0.035 on the test dataset.

1. INTRODUCTION

The United Nations and its member states have set forth the sustainable development goals (SDGs), in which SDG 7 outlines a commitment towards "*ensuring access to affordable, reliable, sustainable, and modern energy for all*" (Sachs, Kroll, Lafortune, Fuller, & Woelm, 2022). Five key targets have been identified towards attaining this goal. Targets 7.1 and 7.2 are of particular interest (*Goal 7: Affordable and clean energy*, 2024):

- Universal access to affordable and clean energy sources prioritizing the transition to renewable energy and energy-efficient technologies by 2030 (Target 7.1).
- Increasing the share of renewable energy in the global energy mix, encouraging the adoption of cleaner and greener alternatives to fossil fuels by 2030 (Target 7.2).

Wind energy, with its meteoric growth in recent years, will play a significant role in contributing towards these targets. For example, the share of wind energy in the global energy mix has increased from 342 TWh in 2010 to 2, 125 TWh in 2022 (International Energy Agency, 2023). With many large-scale wind projects in various stages of development, this trend is expected to continue in the coming years as well.

Wind turbines in a farm are often exposed to complex and harsh operational environments which adversely affects its health conditions and thereby its life expectancy. The average lifetime of wind turbines varies from 20 to 25 years, depending on the design features and operational environment (Adedipe, & Shafiee, 2021; Ziegler, Gonzalez, Rubert, Smolka, & Melero, 2018). During this period, wind turbines undergo gradual degradation in performance owing to the mechanical wear and tear over the years (Hamilton, Millstein, Bolinger, Wiser, & Jeong, 2020; Pan, Hong, Chen, Feng, & Wu, 2021), or the reduction in aerodynamic efficiency due to material erosion over the blade tips (Mathew, Kandukuri, & Omlin, 2022; Ravishankara, Ozdemir, & Weide; Sareen, Sapre, & Selig, 2014). It is estimated that, in Europe, nearly half of the wind turbines in operation will reach their end of designed life by 2030 (Windeurope Asbl/Vzw, 2024). Thus, estimation of the long-term performance of wind turbines in a farm is essential for identifying the possible system degradations over the years and thereby to plan the maintenance strategies and end-of-life decision support.

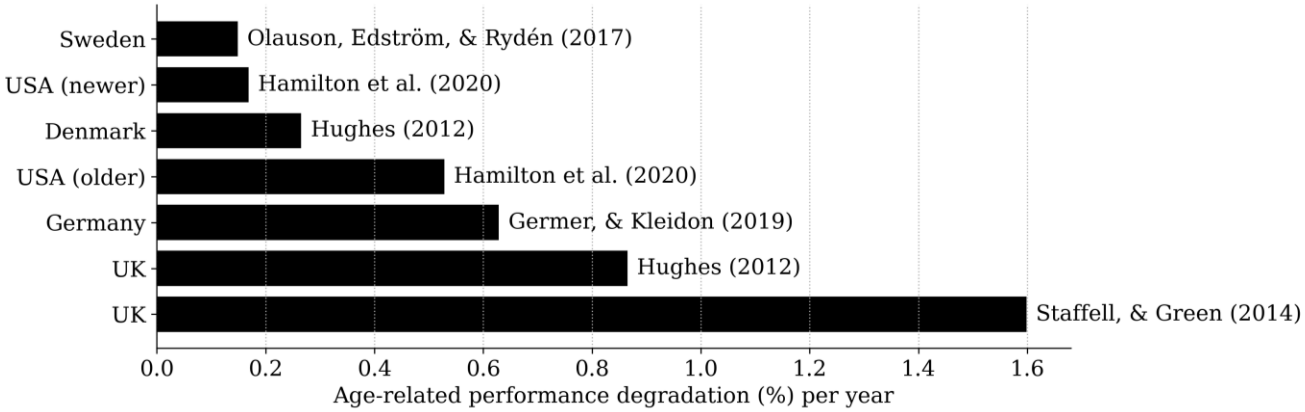


Figure 1. Age-related decline in capacity factor as reported by different studies.

Despite the significance of understanding the health status of wind turbines through its performance degradation during its lifetime, most of the earlier studies on condition monitoring focus solely on component level system reliability and availability (Staffell, & Green, 2014).

Wind turbines have several components integrated within the system and several of such turbines work together with mutual interactions in a wind farm. Hence, an analysis at turbine and farm level would help in giving a wholistic picture of the degradation issues. With the extensive deployment of supervisory control and data acquisition (SCADA) systems, time series performance of wind turbines and farms can be analyzed using data-driven models. Further, the degradation pattern in wind turbines over their life span is highly site-specific in nature (Mathew et al., 2022). Thus, data-driven models can help in estimating the performance degradation in wind turbines accurately and accounting for site-specific factors leading to their degradation. The authors have earlier developed a site-specific degradation estimation model for a wind turbine operating in Norway (Mathew et al., 2022). It was found that the reduction in performance of a wind turbine can be estimated using SCADA data and data-driven models. Further, it was estimated that on average, the performance of the wind turbine under study declined 0.64% every year of its operation. Similar studies have been carried out for turbine-level estimation of performance degradation at Irish, and Italian sites (Astolfi, Byrne, & Castellani, 2021; Byrne, Astolfi, Castellani, & Hewitt, 2020), showing degradation estimates of 8.8% and 1.5% over 12 years of operation. Such wide variation in the performance degradations of wind turbines further strengthens the argument for their site-specific analysis.

At a wind farm level, age-related decline in efficiency is quantified using the plant capacity factor (C_f), which is the ratio of the actual energy produced by the wind farm to the maximum possible energy it could have produced if it were operating at full capacity over the same period. In one of the earliest studies in estimating the wind farm level performance degradation, Hughes (2012) calculated the monthly capacity factor of wind farms operating in the UK and Denmark using 10 years of operational data, which was used to estimate the decline in performance of 13% in the UK and 4% in Denmark over the course of its operation, respectively. Similar results were reported by several studies (Germer, & Kleidon, 2019; Hamilton et al., 2020; Hughes, 2012; Olauson, Edström, & Rydén, 2017; Staffell, & Green, 2014) in the literature as illustrated in Figure 1. In the figure, the age-related decline in performance of wind farms estimated using capacity factor is normalized to per year values as reported in these studies.

These studies help in understanding the age-related performance decline in wind farm level and reiterate the regional and site-specific nature of the degradation phenomenon. However, most of these studies are based on cumulative data from different windfarms collected from public databases. Additionally, they depend on modelling the capacity factor based on meteorological reanalysis data and manufacturer's power curve (MPC) of the wind turbine. Hence, these studies are not based on the data measured from the specific wind farm site under study. The site-specific dynamics play a significant role in the age-related performance degradation of wind turbines, and the performance estimated using MPCs generally differ significantly from field performance of the turbines (Veena, Manuel, Mathew, & Petra, 2020). This could adversely affect the accuracy of these analyses. A more systematic and accurate analysis of the wind farm level performance degradation can be achieved through models based on the site-specific data, derived from the SCADA systems.

In this paper, we propose a deep neural network-based model to estimate the capacity factor of wind farms which can further be used for identifying the age-related performance degradation in wind farms. Apart from using the realistic data derived from SCADA for the site-specific analysis as discussed above, another novelty of the study is the use of convolutional neural network (CNN) model with the soft ordering mechanism. The remainder of the paper is organized

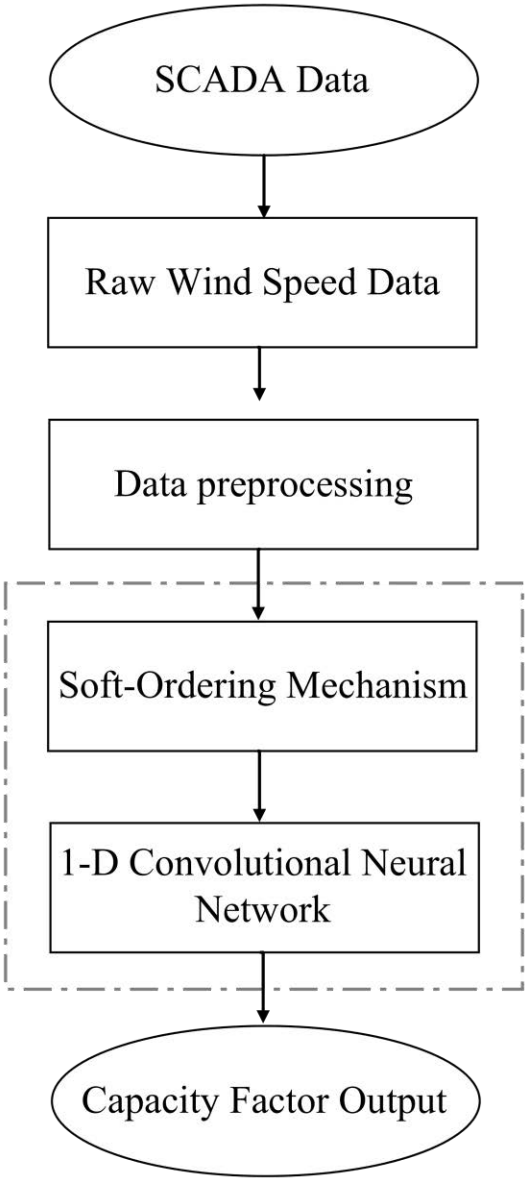


Figure 2. Schematic representation of modelling.

as follows: Section 2 starts by explaining the rationale behind using CNNs. In Section 2.1, the theoretical framework behind CNNs is briefly explained. The soft ordering mechanism employed in order to transform the input data into appropriate inputs to the CNNs is introduced in Section 2.2. Section 2.3 briefly discusses the 1-D CNN architecture and Section 2.4 describes the methodology followed in training, validating, and testing this model. The results from this study are detailed in Section 3 and finally Section 4 concludes this work and traces the next steps in this ongoing study.

2. METHODOLOGY

The performance of wind turbines in a wind farm is significantly influenced by the high spatial and local correlation of wind speed at each of the turbines through site-specific wake effects. But these correlations are further complicated due to the directional and stochastic nature of wind, making it harder for a straightforward analysis. Owing to their capability to extract salient feature representations from data with inherent spatial and local correlations, CNNs are a compelling approach to be explored. The overview of the methodology in estimating the capacity factor is shown in Figure 2.

2.1. Convolutional Neural Networks

The model for estimating the capacity factor in this study is based on CNN. CNNs are inspired by the natural vision in mammals and were popularized by Lecun et al. (1989) particularly for image recognition tasks. Even though the theoretical framework for CNNs predates this work, they used this architecture for automated extraction of features for vision related tasks.

Convolutional layers are the fundamental building blocks in CNN. They serve as the feature extractors exploiting local connectivity, and spatial locality (Kiranyaz et al., 2021; Rawat, & Wang, 2017). In convolutional layers, a learned kernel convolves with the input producing a feature map. The property of local connectivity arises from the fact that each element in the feature map is connected to a local subset of neurons in the previous layer or the input pixels. Spatial locality, on the other hand, is the result of the high correlation between the local subset of input to the convolutional layer. The feature map element at (i, j) in the kth feature map of the lth layer can be calculated as:

$$z_{i,j,k}^{l} = \boldsymbol{w}_k^l \boldsymbol{x}_{i,j}^l + b_k^l \quad (1)$$

where $\boldsymbol{w}_k^l$ and b_k^l are the weight vector and bias term of the k^{th} filter of the l^{th} layer, respectively. $\boldsymbol{x}_{i,j}^l$ is the local subset of input to the convolutional layer centered at (i, j). However, when used for tabular dataset, convolutional layers expect spatial and local correlation between the features. Non-linearity is generally introduced after convolution by using elementwise non-linear activation functions such as rectified linear unit (ReLU). ReLU outputs the input values as such if the input is positive and zero if the input value is negative.

$$a_{i,j,k}^{l} = max\left(0,\, z_{i,j,k}^{l}\right) \quad (2)$$

where $a_{i,j,k}^{l}$ is the activation at position (i, j, k) in layer l after applying ReLU function.

Pooling layers are an optional layer in CNNs which introduce shift-invariance to the feature maps produced by convolutional layers. Shift-invariance is achieved by reducing the resolution of the feature maps through average pooling or max-pooling depending on the task. The average pooling operation is given by:

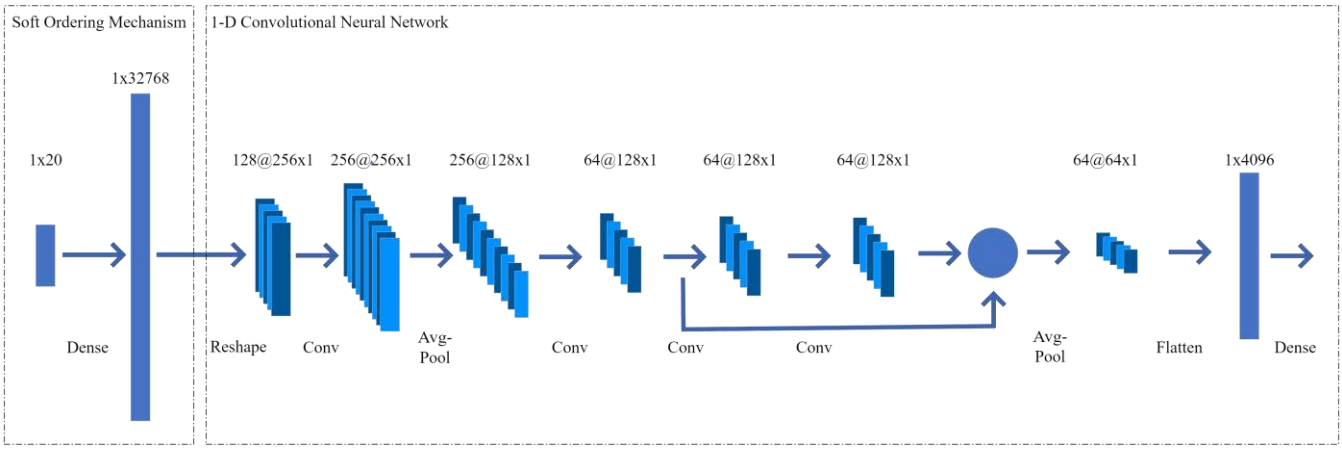


Figure 3. Architecture of the proposed 1-D CNN with soft ordering mechanism.

$$AvgPooling(\boldsymbol{A})_{i,j,k} = {}^{1}/_{N} \sum_{m,n \in R_{i,j}} a^{l-1}_{m,n,k} \qquad (3)$$

where $\boldsymbol{A}$ is the activation map from layer (l-1), N is the number of elements in the pooling window $R_{i,j}$ of dimension (m, n), and $a^{l-1}_{m,n,k}$ is the activations from the pooling layer within the pooling window.

Several convolutional and pooling layers are stacked in a CNN to extract higher level feature representations. Further one or more fully connected (FC) layers are used to achieve higher level reasoning in CNNs (Simonyan, & Zisserman, 2014). The output layer is the final layer that uses task appropriate activation functions (e.g., sigmoid for classification or ReLU for regression).

However, a key challenge is that CNNs are designed to work with data presented in a uniform grid-like structure akin to images. The wind speed input from the wind farm cannot be fed directly to the CNNs assuming each point as a "pixel" in the pseudo-image, as generated from the wind farm layout. This is because the layout of wind turbines in a wind farm is non-uniform, often dictated by the availability of wind and other external factors such as terrain, land use regulations etc.

One solution for the irregular layout of wind farms is to pad the layout with zeros to make it a uniform grid like structure, which results in sparseness in the data. Sparseness in data may result in slowing down of training, reduction in model performance, and loss of spatial resolution.

To overcome these limitations, in this study, we propose a novel application of soft ordering mechanism for CNNs in estimation of the capacity factor. Under this method, the wind data, which is in tabular form, is reshaped into a multi-channel image format. The advantage of this method is that the spatial or sequential relationships of the data are preserved without the need for following a rigid order. This makes the proposed method unique and more suitable for modelling wind farms, which normally have nonrigid geometries. The proposed soft ordering 1-D CNN consists of two parts: a soft-ordering mechanism, and a 1-D CNN.

2.2. Soft ordering mechanism

Soft ordering is a technique to rearrange the data to introduce or preserve spatial or sequential relationships without following a rigid order. In this work, soft ordering is achieved by using an FC layer. The FC layer maps the input features into another higher dimensional feature space. This transformation helps in providing enough pseudo-pixels for the convolutional layers as well as to reorganize the features such that it mimics the spatial or sequential relationships in the data. The FC layer is followed by a non-linear activation function, ReLU in this work, for ensuring that the transformation can effectively learn a non-linear mapping.

Finally, the newly rearranged features are reshaped into multi-channel pseudo-images. Thus, the convolutional layer extracts the features from a rearranged non-linear transformation of the original data, and the model learns to effectively rearrange the features adaptively. Thus, the entire model can be trained in an end-to-end manner without significant preprocessing steps.

The soft ordering mechanism is shown in Figure 3, which takes in the input features and transforms them into non-linear higher-dimensional representations of size 32768. These representations are then reshaped into 128 channels with a signal size of 256 to be fed into the 1-D CNN.

2.3. 1-D CNN Architecture

As opposed to CNNs used for image tasks, where the convolution is applied to a 2-D tensor, a 1-D convolutional layer takes in a single dimensional signal and applies a convolutional kernel of similar dimensionality, typically smaller than the signal. This makes it suitable for applications

like natural language processing, audio signal processing, and time series analysis.

In this work, the representations from the soft ordering mechanism are fed into the 1-D CNN. The 1-D CNN architecture is also shown in Figure 3. The first convolutional layer increases the number of feature channels to 256 while maintaining the size of each feature map at 256 by applying a convolution kernel of size 5. Subsequent adaptive average pooling layer reduces the feature map resolution to 128 x 1. The next three convolution layers apply a kernel of size 3 with a stride of length 1 and output 64 channels of feature maps of size 128. A skip connection is also added from the output of the second convolutional layer to the output of the fourth convolutional layer as shown in Figure 3 to solve the problem of vanishing gradients and hence network degradation (He, Zhang, Ren, & Sun, 2016). A second average pooling layer further reduces the size of the feature maps while ensuring enough receptive fields to facilitate learning. Finally, the output from the average pooling layer is flattened and fed into a fully connected layer which makes the capacity factor estimations. ReLU activation function is used throughout the network to introduce non-linearity except to the outputs of the FC layer in the soft ordering mechanism, where continuously differentiable exponential linear unit (CELU) activation (Barron, 2017) has been used. CELU ensures that non-linearity introduced is smooth and continuous for all values and helps in capturing the negative values effectively avoiding dying ReLU problem (Lu, Shin, Su, & Karniadakis, 2019).

Batch normalization has also been implemented to help the model learn faster and make training more stable by reducing internal covariate shift (Ioffe, & Szegedy, 2015). Further, weights normalization is also implemented to counteract vanishing or exploding gradients and improving generalization by preventing the weights from growing too large or too small (Salimans, & Kingma, 2016).

2.4. Network Training

The model was trained on a wind farm dataset operating at a Norwegian site, by collecting 13 years of operational data. Each of the twenty pitch-controlled wind turbines has a 2 MW rated capacity. The turbines have cut-in, rated, and cut-out velocities of 3 m/s, 18 m/s, and 25 m/s, respectively. The turbines had a rotor diameter of 82.4 m and were installed at a hub height of 70 m. The SCADA data from these turbines had a temporal resolution of 10 minutes (Under the non-disclosure agreement, the data cannot be shared with this paper). The wind speeds and power generated by these turbines were collected from the data and cleaned for missing data and outliers. The initial four years of data from 2007 to 2010 was used to train the model.

The capacity factor of the plant was calculated which served as the target variable and the individual wind speeds served as the features. The data was divided into training, validation, and testing sets in the ratio 3:1:1. Huber Loss was used to calculate the losses for back propagation. Huber Loss is given by:

$$l(y,x) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} 0.5\,(\epsilon)^2, & if\ |\epsilon| < \delta \\ \frac{1}{N}\sum_{n=1}^{N} \delta\,(|\epsilon| - 0.5\delta), & otherwise \end{cases} \quad (4)$$

where $\epsilon = y_n - x_n$, is the residual, δ is the threshold for switching between the δ-scaled L1 and L2 losses, and y_n is the model's estimation of x_n. The advantage of the Huber Loss is that it combines the benefits of both L1 loss (absolute error) and L2 loss (squared error) reducing the penalty for residuals less than the threshold and thereby making the model less sensitive to outliers than L2 loss. The Huber loss is sensitive to the threshold (δ) and was set as two times the standard deviation of the residuals from a basic regression model developed initially using inlier data. Additionally, L1 losses and L2 losses across the training epochs were monitored to ensure that the model's improvement on Huber loss is translated into real world improvement in the estimation of the model performance. Adam optimizer was used in this study for updating the parameters with $\beta_1 = 0.8$ and $\beta_2 = 0.999$. The learning rate (LR) for the optimizer was empirically set to 8×10^{-4}, with an exponential LR decay with $\gamma = 0.9$, meaning the LR would decay after each epoch gradually. This helps in having higher adjustments to the parameters in the beginning and relatively smaller ones towards the end of training. The model was trained over 200 epochs implementing an early stopping mechanism that monitors the validation losses with a patience of 25 to avoid overfitting. Further, L2 regularization was implemented to reduce the chances of overfitting. While dropout layers were investigated for better generalization, it was found that the performance of the model was worse, and convergence was very slow. In the next section, we discuss the results of this experiment in detail.

3. Results and Discussions

The various losses tracked during the training and validation phases are shown in Figure 4: (a) Huber loss, (b) L1 loss, and (c) L2 loss. As expected, the losses are high initially then quickly declining to a more gradual and stable loss condition.

The validation losses in all three of the metrics show high variability across the initial epochs as the model begins to learn from the training data quickly stabilizing showing improvements in generalizability of the model. The best performing model was detected at the 94th epoch with a training and validation loss (Huber loss) of $3.1 \times 10^{-3}$and 1.6×10^{-3}, respectively. The higher training loss observed

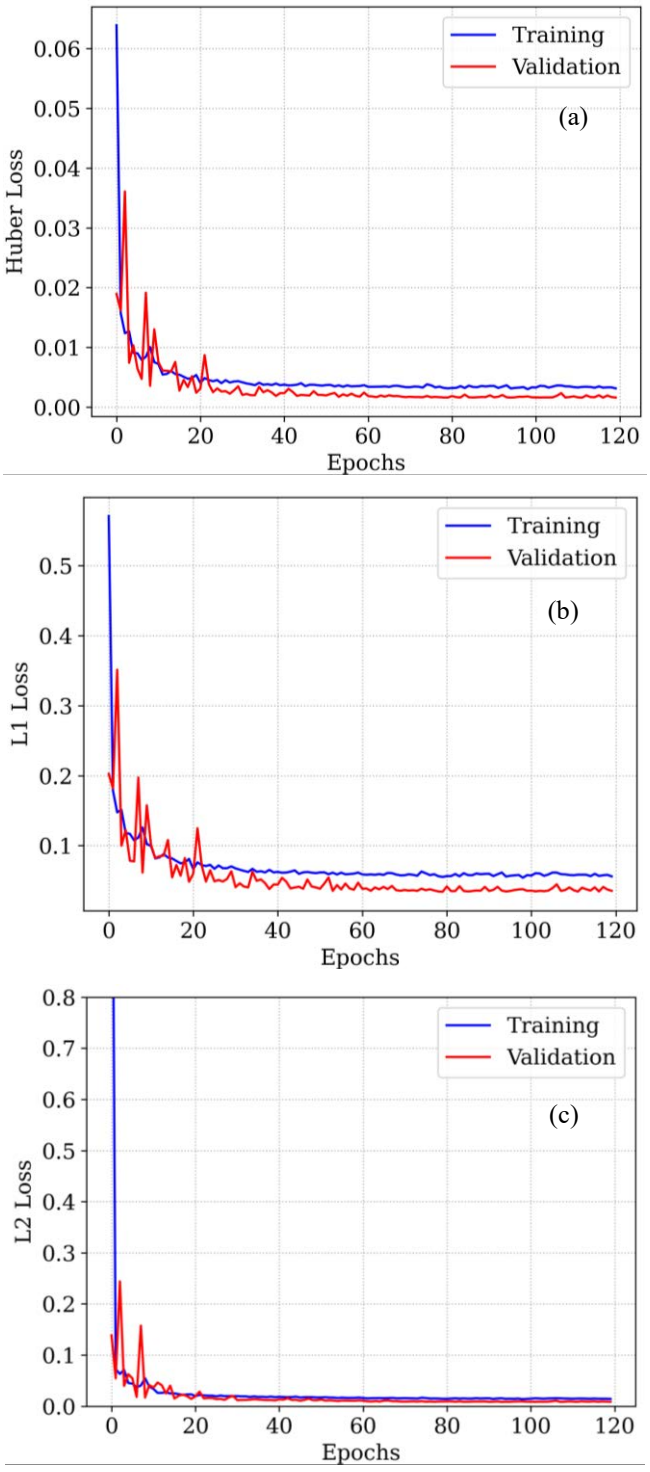


Figure 4. Various losses tracked during training: (a) Huber Loss, (b) L1 Loss, and (c) L2 Loss.

across the epochs are a result of the regularization methods applied during training. The corresponding MAE and mean squared error (MSE) for the training and validation phase can be seen in Table 1.

Table 1. Performance of the best model in training, validation, and test datasets.

Loss	Training	Validation	Test
Huber loss	3.1×10^{-2}	1.6×10^{-3}	1.7×10^{-3}
MAE	5.6×10^{-2}	3.5×10^{-2}	3.5×10^{-2}
MSE	1.4×10^{-2}	8.5×10^{-3}	1.0×10^{-2}

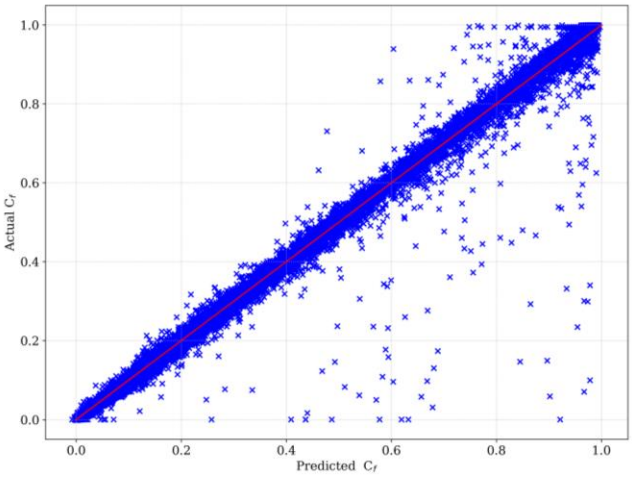


Figure 5. Comparison of the model predicted capacity factor to the measured capacity factor.

The model thus finalized, was tested with test data, which was not used in the training or validation phases to measure the generalizability of the model. Figure 5 shows the performance of the model on the test dataset. The blue scatter indicates the model prediction compared to the calculated values and the distance of these points from the red line indicates the residuals of the prediction model. The training curves (Figure 4) and the comparison in Figure 5, highlight the generalizability of the model to new data and performance of the model on new data, respectively.

The different error metrics in Table 1 quantifies this performance with a slightly higher Huber loss in predicting new datapoints. The normalized root mean squared error in predicting the capacity factor for the test dataset was 0.102. With only 0.363% of the test dataset having a residual value of more than 0.2, the model is found to be effective in capturing the plant capacity factor.

Figure 6 shows the actual and predicted power over different months in a year. It is evident that the predictions and the calculated values are in close agreement with each other. The yearly capacity factor of the farm was calculated as 0.298 against which the model prediction was 0.305. These results further support the argument that the model performs exceptionally well in predicting the wind farm capacity factor.

In previous studies on the age-related performance decline of windfarms, instead of the real data collected from the sites,

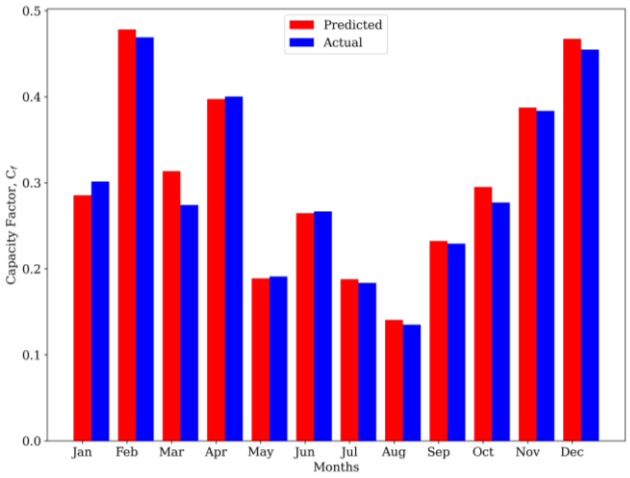


Figure 6. Comparison of actual and predicted power over different months in a year.

the wind estimates from the numerical weather prediction (NWP) models are used for estimating the capacity factors. Though the errors due this approximation are not specified in these studies, obvious differences between the NWP wind predictions and the real velocities available for the turbines could bias the model and thereby adversely affect the reliability of the results. In contrast, real wind measurements are used in the present work which resulted in accurate capacity factor predictions as evident from the low errors values. Similarly, while comparing with some CNN based studies for wind farm performance predictions (Chen et al., 2021; Kazmi, Gorgulu, Cevik, & Baydogan, 2023; Liu et al., 2021), the proposed soft ordering approach could improve the performance of the capacity factor estimations.

4. Conclusion

Wind turbines operating in a farm are exposed to complex operational conditions, causing degradation in their performance over the years of their operation. This age-related performance decline, if quantified at a wind-farm level, could contribute towards making efficient decisions at their end-of-life. As a first step towards this objective, we developed an intelligent algorithm for the estimation of wind farm capacity factor in this paper.

To predict the capacity factor of a wind farm, a 1-dimensional convolutional neural network has been trained exploiting the local connectivity inherent in wind farms. However, to sidestep the irregularity in wind farm layouts, while still using CNNs to model their performance, a soft ordering mechanism is used. The soft ordering mechanism in addition to the 1-D CNN, was able to effectively capture the inherent spatial dynamics in the wind farm as evidenced by the results discussed in the previous section. The model developed in this paper has a normalized root mean squared error of 0.102. This indicates that the errors in the model predictions are approximately 10.2 % of the range of the target values. This indicates that the proposed method could predict the capacity factor of the wind farm with high accuracy. Further, the performance of the model on previously unseen dataset (MAE: 0.035, MSE: 0.010), shows that the model can generalize well to newer data coming from the wind farm even though it was trained on data from earlier.

For developing the proposed model, high quality SCADA data are required, which may limit its applications in farms which do not have such systems in place. Nevertheless, most of the contemporary wind farms have implemented the SCADA systems and with the availability of the required data, the soft ordering 1-D CNN model developed under the study could further be used to estimate the age-related performance degradation in wind farms. This will be demonstrated by the authors through their ongoing research where logs on the turbine maintenance will also be considered.

Acknowledgements

This research work has been funded by Analytics for asset Integrity Management of Windfarms (AIMWind), under grant no. 312486, from Research Council of Norway (RCN).

AIMWind is collaborative research from University of Agder, Norwegian Research Center (NORCE), and TU Delft, with DNV and Origo Solutions as advisory partners.

References

Adedipe, T., & Shafiee, M. (2021). An economic assessment framework for decommissioning of offshore wind farms using a cost breakdown structure. *The international journal of life cycle assessment*, *26*(2), 344-370. https://doi.org/10.1007/s11367-020-01793-x

Astolfi, D., Byrne, R., & Castellani, F. (2021). Estimation of the performance aging of the vestas v52 wind turbine through comparative test case analysis. *Energies*, *14*(4), 915. https://www.mdpi.com/1996-1073/14/4/915

Barron, J. T. (2017). Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*. https://doi.org/10.48550/arXiv.1704.07483

Byrne, R., Astolfi, D., Castellani, F., & Hewitt, N. J. (2020). A study of wind turbine performance decline with age through operation data analysis. *Energies*, *13*(8), 2086. https://www.mdpi.com/1996-1073/13/8/2086

Chen, X., Zhang, X., Dong, M., Huang, L., Guo, Y., & He, S. (2021). Deep learning-based prediction of wind power for multi-turbines in a wind farm. Frontiers in Energy Research, 9, 723775. https://doi.org/10.3389/fenrg.2021.723775

Germer, S., & Kleidon, A. (2019). Have wind turbines in Germany generated electricity as would be expected from the prevailing wind conditions in 2000-2014? *PLoS One*, *14*(2), e0211028. https://doi.org/10.1371/journal.pone.0211028

Goal 7: Affordable and clean energy. (2024). The Global Goals. Retrieved 13/03/2024 from

https://www.globalgoals.org/goals/7-affordable-and-clean-energy/

Hamilton, S. D., Millstein, D., Bolinger, M., Wiser, R., & Jeong, S. (2020). How does wind project performance change with age in the united states? *Joule*, *4*(5), 1004-1020. https://doi.org/https://doi.org/10.1016/j.joule.2020.04.005

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 27-30 June 2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778, arXiv: 1512.03385. https://doi.org/10.48550/arXiv.1512.03385

Hughes, G. (2012). The performance of wind farms in the United Kingdom and Denmark. *Renewable Energy Foundation*, 48.

International Energy Agency. (2023). *World energy outlook 2023* (World Energy Outlook, Issue. https://www.iea.org/reports/world-energy-outlook-2023

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning, 448-456. https://doi.org/10.48550/arXiv.1502.03167

Kazmi, S., Gorgulu, B., Cevik, M., & Baydogan, M. G. (2023). A concurrent cnn-rnn approach for multi-step wind power forecasting. arXiv preprint arXiv:2301.00819. https://doi.org/10.48550/arXiv.2301.00819

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398. https://doi.org/https://doi.org/10.1016/j.ymssp.2020.107398

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

Liu, T., Huang, Z., Tian, L., Zhu, Y., Wang, H., & Feng, S. (2021). Enhancing wind turbine power forecast via convolutional neural network. *Electronics*, *10*(3), 261. https://doi.org/10.3390/electronics10030261

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*. https://doi.org/10.48550/arXiv.1903.06733

Mathew, M. S., Kandukuri, S. T., & Omlin, C. W. P. (2022). Estimation of wind turbine performance degradation with deep neural networks. 7th European Conference of the Prognostics and Health Management Society 2022, 7(1), 351-359. https://doi.org/10.36001/phme.2022.v7i1.3328

Olauson, J., Edström, P., & Rydén, J. (2017). Wind turbine performance decline in Sweden. *Wind Energy*, *20*(12), 2049-2053. https://doi.org/https://doi.org/10.1002/we.2132

Pan, Y., Hong, R., Chen, J., Feng, J., & Wu, W. (2021). Performance degradation assessment of wind turbine gearbox based on maximum mean discrepancy and multi-sensor transfer learning. *Structural Health Monitoring*, *20*(1), 118-138. https://doi.org/10.1177/1475921720919073

Ravishankara, A. K., Ozdemir, H., & Weide, E. v. d. Effect of leading edge erosion on wind turbine rotor aerodynamics. In *Aiaa scitech 2022 forum*. https://doi.org/10.2514/6.2022-0276

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352-2449. https://doi.org/10.1162/neco_a_00990

Sachs, J. D., Kroll, C., Lafortune, G., Fuller, G., & Woelm, F. (2022). *Sustainable development report 2022*. Cambridge University Press. https://doi.org/10.1017/9781009210058

Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, *29*. https://doi.org/10.48550/arXiv.1602.07868

Sareen, A., Sapre, C. A., & Selig, M. S. (2014). Effects of leading edge erosion on wind turbine blade performance. *Wind Energy*, *17*(10), 1531-1542. https://doi.org/https://doi.org/10.1002/we.1649

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Staffell, I., & Green, R. (2014). How does wind farm performance decline with age? *Renewable Energy*, *66*, 775-786. https://doi.org/https://doi.org/10.1016/j.renene.2013.10.041

Veena, R., Manuel, S. M., Mathew, S., & Petra, I. (2020). Parametric models for predicting the performance of wind turbines. *Materials Today: Proceedings*, *24*, 1795-1803. https://doi.org/https://doi.org/10.1016/j.matpr.2020.03.604

WindEurope asbl/vzw. (2024). *Wind energy today | windeurope*. Retrieved 13/03/2024 from https://windeurope.org/about-wind/wind-energy-today/

Ziegler, L., Gonzalez, E., Rubert, T., Smolka, U., & Melero, J. J. (2018). Lifetime extension of onshore wind turbines: A review covering germany, Spain, Denmark, and the UK. *Renewable and Sustainable Energy Reviews*, *82*, 1261-1271. https://doi.org/https://doi.org/10.1016/j.rser.2017.09.100

BIOGRAPHIES

Manuel S. Mathew is a PhD Research Fellow at the Information and Communication Technology department at the University of Agder, Norway. His interest is in the application of artificial intelligence in renewable energy systems particularly focusing on prognostics for wind farms. He completed his master's degree in Renewable Energy in 2021 from the University of Agder. In addition, he also holds a master's degree in systems engineering by research from the University of Brunei Darussalam. He did his bachelor's degree in electrical and electronics engineering from the Mahatma Gandhi University, India.

Surya Teja Kandukuri is a Senior Scientist at NORCE. He holds a part-time position as a Researcher at University of Agder, Grimstad, Norway. He obtained his PhD in condition monitoring from the University of Agder in 2018. He has over 12 years of experience in industrial research within aerospace, energy, marine and oil & gas sectors, developing condition monitoring solutions for high-value assets. He received his master's degree in systems and control engineering from TU Delft, The Netherlands, in 2006 and bachelor's in electrical engineering from Nagarjuna University in India in 2003.

Christian W. Omlin has been a professor of Artificial Intelligence at the University of Agder since 2018. He has previously taught at the University of South Africa, University of the Witwatersrand, Middle East Technical University, University of the South Pacific, University of the Western Cape, and Stellenbosch University. His expertise is in deep learning with a focus on applications ranging from safety to security, industrial monitoring, renewable energy, banking, sign language translation, healthcare, bio conservation, and astronomy. He is particularly interested in the balance between the desire for autonomy using AI technologies and the necessity for accountability through AI imperatives such as explainability, privacy, security, ethics, and artificial morality for society's ultimate trust in and acceptance of AI. He received his Ph.D. from Rensselaer Polytechnic Institute and his MEng from the Swiss Federal Institute of Technology, Zurich, in 1995 and 1987, respectively.

State-of-Charge and State-of-Health Estimation for Li-Ion Batteries of Hybrid Electric Vehicles under Deep Degradation

Min Young Yoo[1], Hyun Joon Lee[1], Woosuk Sung[2], Jae Sung Heo[3] and Joo-Ho Choi[1*]

[1]*Korea Aerospace University, Goyang-si, Gyeonggi-do, 10540, Republic of Korea*
myyoo@kau.kr
moon0601jk@kau.kr
**Corresponding author: jhchoi@kau.ac.kr*

[2]*Chosun University, Gwangju, 61452, Republic of Korea*
wsung@chosun.ac.kr

[3]*Korea Aerospace Research Institute, Daejeon, 34133, Republic of Korea*
jshuh@kari.re.kr

ABSTRACT

In recent industry, hybrid vehicles are gaining more recognition as a practical means for future transportation due to the longer distance, reduced charging time, and less charging stations dependency. The batteries in the hybrid vehicles, however, undergo more complex operation of charge depleting and sustaining modes alternately, which may need more accurate battery state estimation. In this study, a model based method is explored for the Li-ion batteries in the hybrid electric vehicles to estimate State-of Charge (SOC) and State-of-Health (SOH) accurately. While there have been widespread studies for this topic in the batteries research, not many are found that have investigated hybrid operation modes. Also the estimations are mostly limited to normal batteries or shallow degradation with the SOH higher than 90%. In this study, an algorithm based on the dual extended Kalman filter (DEKF) and enhanced self-correcting (ESC) model is developed for the simultaneous estimation of the SOC and SOH. Degradation data for plug-in hybrid vehicle (PHEV) are taken for the study, which undergo the deep degradation of 30%. In order to maintain the accuracy such that the root mean square error (RMSE) of the SOC is within 5% over the entire degradation cycles, two practical methods are proposed: First, the SOH is estimated separately during the battery charging, and is used as a constant in the SOC estimation in the discharging cycles. Second, battery modeling is conducted and the parameters are reset in every intermittent cycles at which the SOH is reduced by 10% initially and by 5% thereafter.

1. INTRODUCTION

Lithium-ion batteries have been applied extensively in various fields, including portable electronic devices, road transportation, and power supply systems, expecting their future role in energy sustainability (Zubi et al., 2018). As battery-powered vehicles such as pure electric and hybrid electric vehicles gain popularity, the development of battery management systems (BMS) estimating the state-of-charge (SOC) and state-of-health (SOH) of the batteries becomes crucial to ensure reliable and efficient battery operation (Mishra et al., 2021). In the BMS research, most SOC and SOH estimators have been developed for pure electric vehicles that primarily operate in charge-depleting (CD) mode. However, there is an increasing demand for hybrid vehicles that can handle higher loads and longer distances, which involves switching between CD and charge-sustaining (CS) mode during the operation. This can make the SOC estimation more difficult than those in the CD mode alone. Therefore, SOC and SOH estimation under combined mode is necessary for improved accuracy (Yoo et al., 2023).

Fundamentally, it is impossible to measure the SOC and SOH of batteries directly, thus methods are designed for estimating them based on measurable data such as current, voltage and temperature. Among the many achievements, Kalman Filter-based algorithms, which belong to the model-based approach, have proven their effectiveness and account for more than half of the SOC estimation methods (Shrivastava et al., 2019). Nevertheless, it is a challenge to estimate SOC for degraded batteries, which requires the SOH estimation as well (Hannan et al., 2017). Investigations into the simultaneous estimation of SOC and SOH, in view of both the effectiveness and efficiency, have remained relatively

Table. 1. Literature using model-based SOC and SOH estimation with aged battery data.

Author	Year	Battery Model	Methods (SOC-SOH)	Estimation Factor	Results of Estimation	Level of Degradation
R. Xiong	2014	1 RC model	EKF-EKF (multi-scale)	SOC, All Parameters	SOC, Capacity	100%, 82.6%, 82.1% and 72.1%
N. Wassiliadis	2018	2 RC model	EKF-EKF	SOC, Capacity, Resistance	SOC, Capacity, Resistance	100%, 97%, 85%, 78% and 49%
J. Wu	2019	2 RC model	AEKF-KF	SOC, Resistance	SOC	96.5%, 93.9%, and 92.4%
X Hu	2018	fractional second-order model	EKF-EKF	SOC, Capacity, Resistance	SOC, Capacity, Resistance	86.1%, 81.7% and 74.5%
L. Ma.	2022	fractional second-order model	MIUKF-UKF (multi-scale)	SOC, Capacity, Resistance	SOC, Capacity	98.1%, 94.7%, and 91.5%

insufficient (Wang et al., 2021). Even in the simultaneous estimation of SOC and SOH, substantial portion have targeted normal batteries (Campestrini et al., 2016; H. Guo et al., 2017; R. Guo & Shen, 2022; Hossain et al., 2022; C. Hu et al., 2012; Lee et al., 2008; Plett, 2004, 2006; Shrivastava et al., 2022; Ye et al., 2023; Zhang et al., 2016)

Model-based SOC and SOH estimation for aged battery can be categorized into two groups. The first involves updating the model's parameters to account for battery aging (Li et al., 2019; Sepasi et al., 2014; Shrivastava et al., 2019; Xu et al., 2022). While it is possible to update the model through optimization, using which the SOC is estimated, it comes at the cost of high computational burden. Additionally, there is a challenge in determining an appropriate updating period. The second approach involves co-estimation of the states and parameters of a battery model (X. Hu et al., 2018; Ma et al., 2022; Wassiliadis et al., 2018; Wu et al., 2019; Xiong et al., 2014). While this can be achieved using the dual filter algorithms (Yoo et al., 2023), it presents a significant challenge due to the substantial number of parameters in the model. This is further compounded by the fact that the only directly measurable output is the voltage under the given currents. Consequently, only a few parameters such as the capacity, i.e., the SOH, and internal resistance are estimated, while the others are held at fixed values. However, this approach may result in a less accurate model of aged battery.

Upon the survey of relevant literature, it follows that the approaches on the co-estimation of SOC and SOH by the dual filters need a comprehensive discussion in various aspects: battery model, type of filters, specific settings of these filters, initial values of state and parameters, and the level of degradation. Regarding the battery model, used models are Thevenin model with first-order (1RC) (Xiong et al., 2014) or second-order (2 RC) (Wassiliadis et al., 2018; Wu et al., 2019), or fractional second-order model (X. Hu et al., 2018; Ma et al., 2022). While the extended Kalman filter (EKF) is usually used, other filters such as adaptive extended Kalman filter (AEKF) or unscented Kalman filter (UKF) have sometimes been used, and there is a case where a dual filter has different time intervals considering the characteristics of state and parameter. Regarding the settings of filter (such as error, noise and measurement covariance), many did not specify values and conditions, except (X. Hu et al., 2018; Wassiliadis et al., 2018).This may make the results less trusted in terms of practical application. In view of the degradation levels, only one paper (Wassiliadis et al., 2018) has explored capacity fade over 50%, but the results are given without confidence intervals. As a result, despite the abundance of literature, these limitations pose challenges in adopting a practical approach to BMS development. Table 1 summarizes the representative papers in terms of model, methods, results of estimation, and level of degradation.

This study presents a more practical methodology to co-estimate the SOC and SOH by the dual Kalman filter for the batteries undergoing hybrid operations. Two key insights are applied for this research. First, it is observed that the co-estimation of SOC and SOH may yield inaccurate results due to the poor observability of the capacity. To mitigate this, a

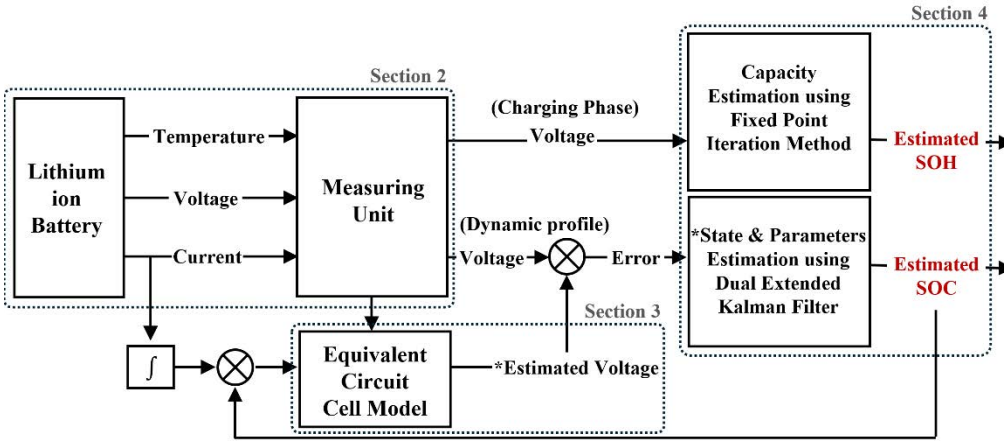


Figure 1. An overview of the core methodology

practical solution is developed by separating the SOC and SOH estimation, namely, estimating the SOC using the Dual Extended Kalman Filter (DEKF) in the discharging phase while estimating the SOH using the Fixed-Point Iteration Method (FPIM) in the charging phase. The reason is that the capacity estimation during the charging process is generally standardized and exhibit less dynamics as opposed to the discharging process. Second, when the batteries degrade by more than 10% in capacity, even the performance of this approach falls below acceptable level. Therefore, remedial action is applied by updating the parameters of battery model periodically.

The approach is validated by utilizing thirty battery cell datasets. These datasets are obtained through tests conducted under three distinct dynamic profiles representing plug-in hybrid electric vehicles (PHEVs), captured at ten points throughout the cycles ranging from 0% to 30% of capacity fade. Section 2 outlines the experimental method to measure temperature, voltage and current, and three types of dynamic profiles in charging phase. In Section 3, two battery models: Thevenin and Enhanced Self-Correcting (ESC) are addressed, which is to estimate voltage from the measured data. Section 4 explains the procedure of SOC and SOH estimation by the DEKF and FPIM respectively. An overview of the key methodology proposed in this paper can be found in Figure 1. Finally, key findings are summarized in Section 5, providing comprehensive insights and limitations.

2. BATTERY CELL TEST

In this study, the same battery cell and equipment described in the literature (Yoo et al., 2023) are used for the test, which is a Samsung SDI, 18650-35E lithium-ion battery cell with nominal capacity of 3.5 Ah and a nominal voltage of 3.7 V. The battery is operated by a DC electronic load (Kikusui, PLZ1004W), a DC power supply (Kikusui, PWR800L) and a charge-discharge system controller (Kikusui, PFX2512) as shown in Figure 2. The test profiles are divided into dynamic test and aging test. The dynamic profiles comprise of three scenarios of Plug-in Hybrid Electric Vehicles (PHEV) as shown in Figure 3: City, Highway, and High-speed. City and Highway profiles consist of charge-depletion (CD) mode and charge-sustaining (CS) mode, while High-speed has CD mode only.

Figure 2. Experimental setup

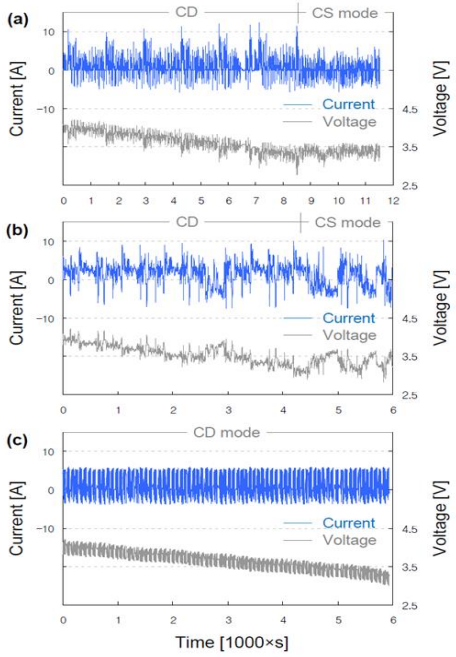


Figure 3. Voltage response to the current profiles adopted: (a) City, (b) Highway, and (c) High-speed

Aging test is performed to acquire aged battery data, by repeating charging and discharging cycles up to the capacity degradation of 30% as shown in Figure 4. Once the battery cell exhibits a noteworthy degree of capacity fade, three different dynamic profiles are applied. Before proceeding the aging test, static test is conducted to obtain the relation between the open circuit voltage (OCV) and SOC as shown in Figure 5. In the aging test, 2100 cycles are used to make capacity fade of 30%. Dynamic tests are conducted at 10 time points with the intervals of every 100 cycles during the period from the initial to the 300th cycles, and with the intervals of every 300 cycles from the 300th to the end of the cycles, which is depicted in Figure 6. The failure threshold for SOH is given at the 80% of initial capacity as shown in the dotted horizontal line in the figure. The capacity decreases mostly in linear fashion, except from 1200 to 1500 cycles where it is constant. The dynamic test data at each cycle, which are 0, 300, 1200 and 2100 cycles, are presented in Figure 7. For each cycle, three dynamic profiles are applied with the initial SOC set at approximately 0.9 (90%). As the degradation proceeds, each profile exhibits an abrupt termination because the cutoff voltage is reached earlier, indicating that the capacity has faded.

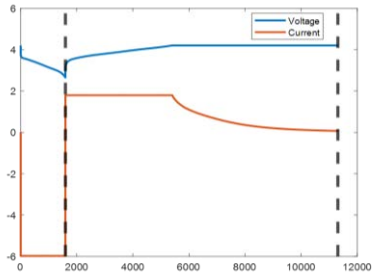


Figure 4. Aging test profiles

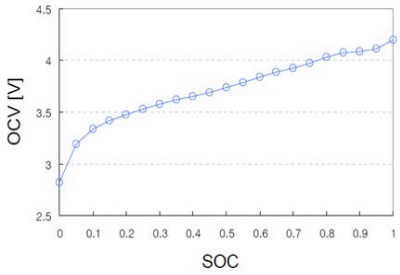


Figure 5. OCV-SOC relationship

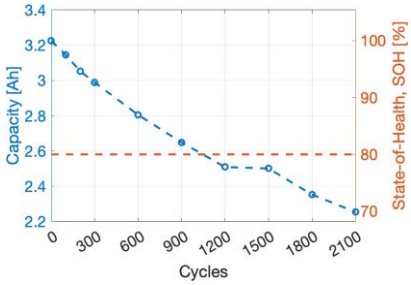
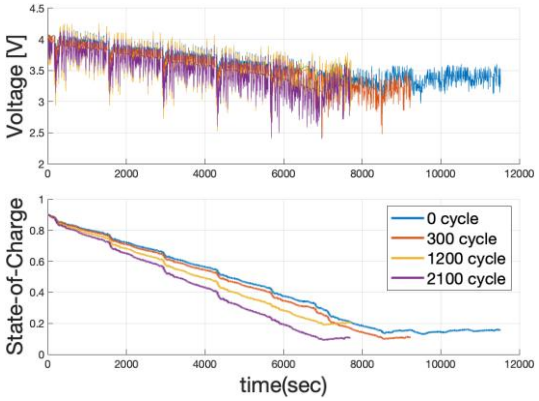
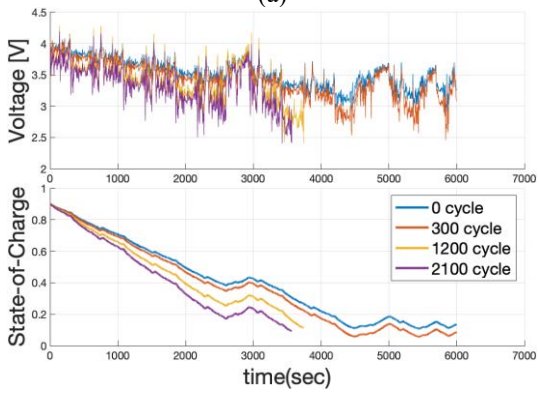


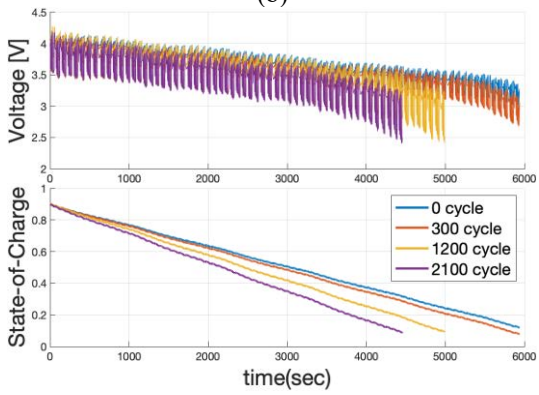
Figure 6. Capacity degradation in the aging test



(a)



(b)



(c)

Figure 7. Voltage response to the current profiles adopted at each cycle: (a) City, (b) Highway, and (c) High-speed

3. BATTERY MODEL

In this study, Thevenin and Enhanced Self-Correcting (ESC) models are reviewed to select a suitable battery model based on the gathered test data. The Thevenin model, one of the most widely utilized equivalent circuit models (ECMs) for model-based estimation of batteries, describes battery behavior by accounting for voltage drop through a resistor element and time-varying polarization voltages through one or more parallel resistor-capacitor (RC) elements.

The ESC model, proposed by Plett (2015), extends the Thevenin model by incorporating a hysteresis term to describe the hysteresis voltage of batteries with empirical modeling. Its configuration is shown in Figure 8, where v_T is terminal voltage, v_{oc} is open-circuit voltage, z is SOC, i is current (current bias ib is ignored in this study) flowing through R_0 (ohmic resistance), i_R is the current flowing through R_j and C_j (polarization resistance and capacitance), h is hysteresis, M is maximum hysteresis voltage, M_0 is instantaneous hysteresis voltage, and s is sign function of i.

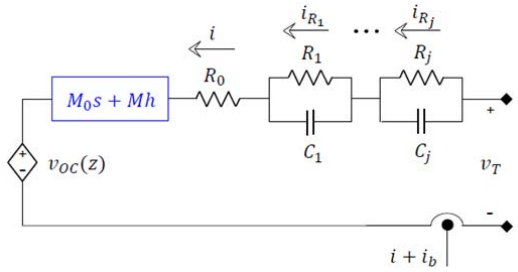
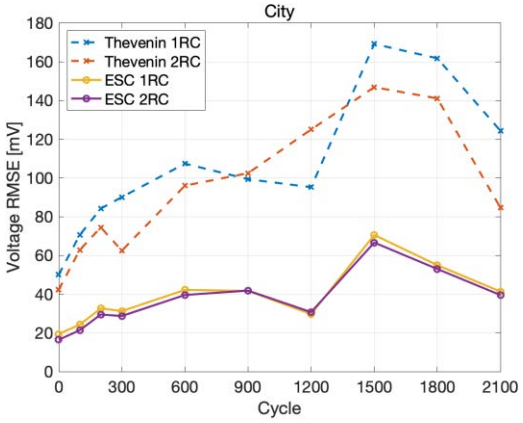
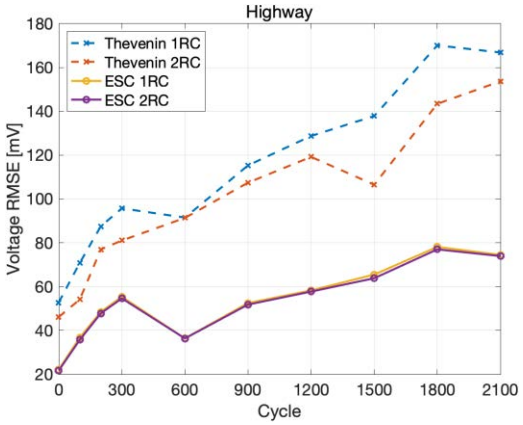


Figure 8. Circuit schematic for the ESC model which is the same as the Thevenin model except the addition of hysteresis voltages (in blue)

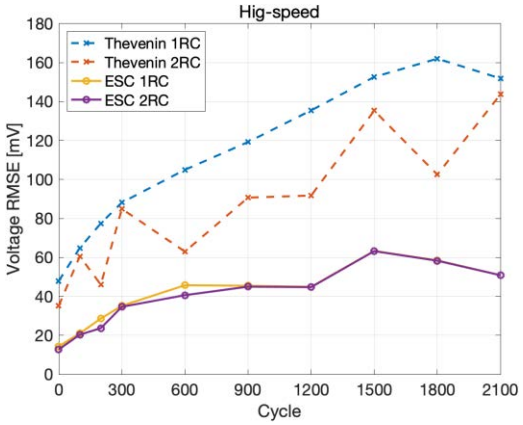
This study considers cases with 1 or 2 RC pairs for both the Thevenin and ESC models, denoted as Thevenin 1RC/2RC, and ESC 1RC/2RC. In order to estimate the model parameters, method by Plett, (2015) and Yoo et al., (2023) is employed, using the dynamic data for each cycle. Consequently, the models enable calculation of terminal voltage under the given current. The performance of each model is summarized in Figure 9 by the root mean square error (RMSE) between measured and estimated voltage. It is noteworthy that the ESC model outperforms the Thevenin model consistently for all the collected data. This superiority becomes more remarkable as the battery is aged. It is found that the hysteresis term in the model is useful to describe aging of the battery. Another observation is that the incorporation of additional RC pairs primarily yields a positive effect in the case of the Thevenin models, whereas it does not in the ESC model. This is from the fact that the hysteresis term diminishes the relative influence of additional RC pairs in the ESC model. This observation supports that adding the hysteresis term is better than adding the number of RC pairs. Therefore, the ESC 1RC model is chosen in this study. However, it is important to note that even the performance of the ESC model experiences accuracy loss in the aged batteries, which means that the model error increase is inevitable as the battery ages.



(a)



(b)



(c)

Figure 9. Modeling performance of each model: (a) City, (b) Highway, and (c) High-speed

4. SOC AND SOH ESTIMATION

In this section, the procedure of SOC and SOH estimation by the DEKF and FPIM are outlined along with the corresponding result. The initial step involves the application of a general DEKF as reviewed in Section 1, based on the

ESC 1 RC model as determined in Section 3. Subsequently, the inherent weak observability of capacity in the DEKF algorithm is identified. To address this, the FPIM, a capacity estimation technique, is integrated into the DEKF framework. However, despite this integration, there exsists a decline in estimation performance for deep degraded battery cells. Therefore, remedial solution is proposed: an initial parameter estimation update is recommended after a 10% capacity loss to enhance estimation performance. Estimation results are summarized in each step of this process. In summary, the contribution of this study is to improve and validate methodologies to estimate SOC and SOH for test datasets with various dynamic profiles, including hybrid operation modes, and various degradation levels, up to 30% of capacity loss of battery cells.

The dual extended Kalman filter (DEKF), one of the approaches to generalize the extended Kalman filter (EKF) for simultaneous estimation of state and parameters, comprises of two filters: one for estimating the state and the other for estimating the parameters (Plett, 2005, 2015). Each filter executes the steps, and they are linked by exchanging information during the time update sequence. In this research, we propose a hybrid approach incorporating the DEKF and the capacity estimation technique to overcome the inherent weak observability of the capacity in the DEKF algorithm (Wassiliadis et al., 2018). The capacity estimation is implemented by the fixed-point iteration method (FPIM) during battery charging, and the estimated capacity value is used as a known value in the DEKF during the battery discharging (Sung & Lee, 2018).

At first, the DEKF is applied for the co-estimation of SOC and SOH, and the results are evaluated by the RMSE in the case of SOC and the last estimated value in the case of SOH in each dynamic profile, whose true values were obtained through coulomb counting of current during the conducted profiles and capacity testing after the profiles, respectively. Then, our proposed approach is applied, where the SOH estimation is separated from the DEKF and is made by the FPIM during the charging cycle.

Estimation Results of SOC and SOH using DEKF

By applying the DEKF based on the ESC 1RC Model to all 30 datasets, estimation results are obtained for SOC and SOH. Regarding the SOC, the RMSE for each profile and cycle is illustrated in Figure 10. It was observed that prior to 600 cycles, the estimation performance exhibits RMSE of less than 3% for all datasets. However, beyond 600 cycles, a significant degradation in estimation performance become evident. As the degradation progresses, it can be observed that the error in the initial SOC gradually increases.

Involving OCV-SOC tests at specific time intervals and subsequently updating the battery model can mitigate inaccuracies in estimation. Nevertheless, static tests for acquiring OCV-SOC lookup tables are time-consuming, and selecting suitable test time points poses a challenge.

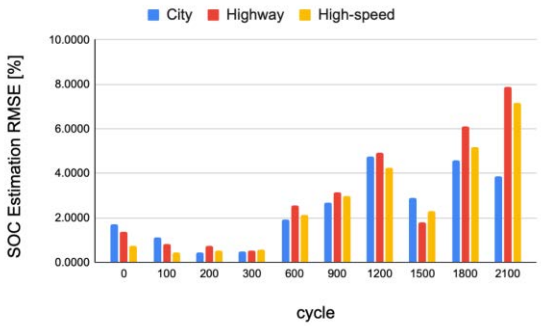


Figure 10. Estimation results of SOC using DEKF

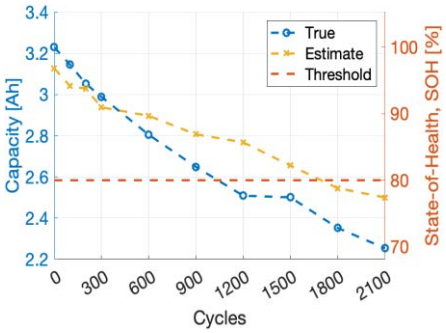
Table. 2. Initial SOC estimation of each dataset

Cycle	City	Highway	High Speed
0	0.88	0.88	0.88
100	0.88	0.88	0.88
200	0.87	0.88	0.88
300	0.87	0.87	0.87
600	0.85	0.85	0.85
900	0.84	0.83	0.83
1200	0.81	0.82	0.82
1500	0.81	0.82	0.83
1800	0.79	0.80	0.80
2100	0.79	0.78	0.78

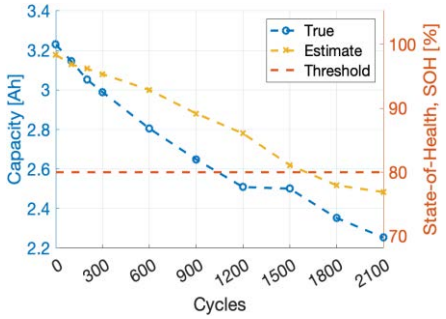
In this paper, a methodology is explored to mitigate inaccuracies in estimation without additional static tests. Hence, the inaccurate initial SOC estimation due to the battery aging remains as an inherent error in estimation without updating the battery model.

In the SOH estimation, it was observed that, except for the early degradation stage, the capacity generally does not satisfy the acceptable level of performance. Figure 11 depict the results of SOH estimation as the battery ages for each profile. It is observed that the capacity has low observability, and its estimation depends on the specific profile. This may be due to the fundamental issues in estimating various states and parameters based on limited measurement data. Meanwhile, for internal resistance, overall estimation performance was found to be superior when compared to the reference values. It is noted that the internal resistance has high observability compared to the capacity. Thresholds for failure based on each SOH were established at 80% of the initial capacity and twice the initial value for the internal resistance. Reference values of internal resistance were derived from battery modeling results rather than obtained through power tests. In this investigation, capacity was regarded as the indicator of SOH since the true capacity was measured at each time point. Consequently, the failure point of the battery is estimated to occur around 1,000 cycles, coinciding with the battery's capacity dropping below 80% of

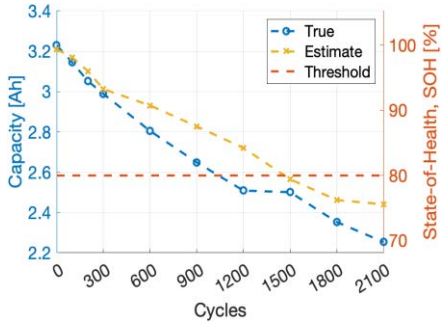
its initial capacity, as shown Figure 12. This failure point couldn't be accurately predicted due to the low observability of capacity in this estimation algorithm.



(a)

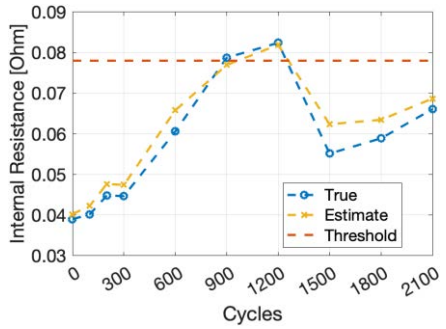


(b)

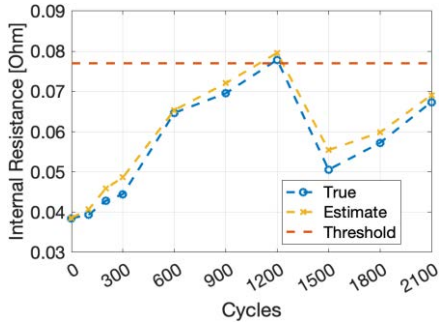


(c)

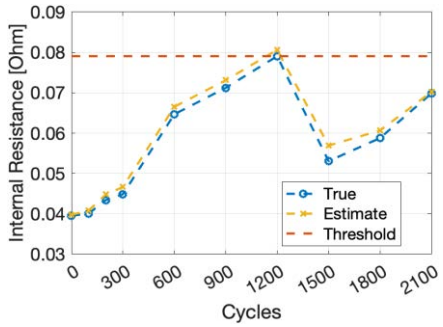
Figure 11. Capacity estimation results of PHEV datasets (a) City, (b) Highway and (c) High-speed



(a)



(b)



(c)

Figure 12. Resistance estimation results of PHEV datasets (a) City, (b) Highway and (c) High-speed

Estimation Results of SOC and SOH using DEKF and FPIM

SOH Estimation

To overcome the low performance in capacity estimation, we have applied capacity estimation techniques separately using the charging profile data. Among various techniques, fixed-point iteration method (FPIM) was selected to estimate capacity during charging. Since there is only one type of profile in the charging, we can obtain a single capacity estimation result, as shown in Figure 13.

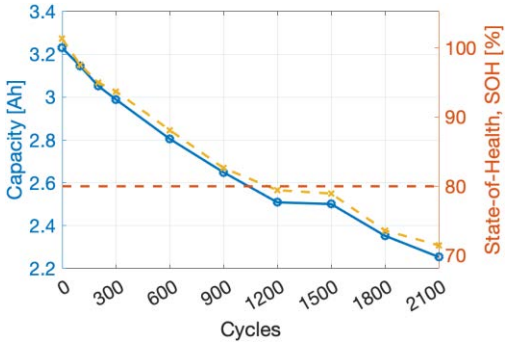


Figure 13. Results of the SOH estimation.

The estimation error was confirmed to be within 3%, indicating a significant improvement in estimation performance compared to the DEKF methodology, which estimates simultaneously from each profile.

SOC Estimation

Utilizing the estimated capacity during charging as a fixed value for SOC estimation using DEKF, the performance of SOC estimation is depicted in Figure 14. Unfortunately, despite improvements in capacity estimation performance, there is no corresponding enhancement observed in SOC estimation. It is observed that the performance significantly deteriorates with RMSE exceeding 5% after 600 cycles.

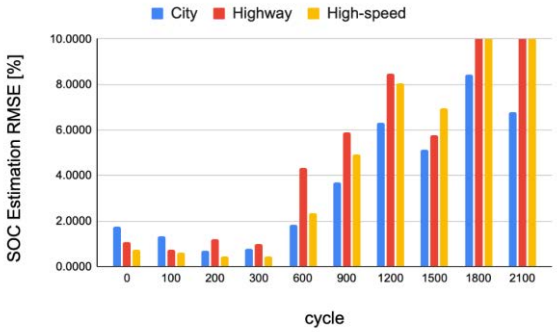


Figure 14. Estimation results of SOC using DEKF and FPIM

The everlasting decrease in the performance of SOC estimation as battery ages, despite the capacity being so close to the true value, can be attributed to several factors. These encompass the initial SOC estimation error and inaccuracies in model assumptions, as previously mentioned. Additionally, nonlinearities in the battery behavior and limitations in the estimation algorithms employed contribute to this phenomenon. Moreover, the interaction among these factors can exacerbate the complexity of the estimation process, further impeding the accurate SOC estimation. In the battery modeling of Figure 9, it has been observed that the ESC 1RC demonstrates sufficient capability to simulate the behavior of aged batteries. Consequently, there is an expectation that the SOC estimation would perform well if the parameters were estimated to optimal values. However, the challenge arises when attempting to simultaneously estimate the states and parameters to their optimal values within the filter algorithm, especially based on the limited measurement data. To address them, the initial parameter estimation values were set as the last parameter estimation values from the former dataset for each profile. This approach aimed to leverage the previous dataset's knowledge and fine-tune the initial parameter values for improved estimation accuracy in subsequent cycles. By initializing the parameters with values derived from the previous dataset, it was expected that the model could benefit from the accumulated insights and trends observed in earlier profiles, thereby enhancing the robustness and reliability of the estimation process. However, while it has been noted that internal resistance exhibits high observability in the estimation algorithm, the majority of parameters display low observability. Furthermore, unlike capacity or internal resistance, there is no discernible trend for each parameter during degradation. This lack of observable trends makes the efforts useless in the estimation process.

Estimation Results of SOC using DEKF, FPIM and initial parameter estimation update

SOC Estimation

To address the issue of deteriorating SOC estimation performance despite the improvement in SOH estimation through separate estimation, a method was applied to conduct the battery model at specific time points and reset the initial parameter estimation values.

Based on the degradation threshold of 20% capacity loss as the failure point for SOH, the battery modeling was conducted using dynamic profiles at intervals of 10% capacity loss initially, and 5% capacity loss subsequently. This method involved conducting the battery model at 600, 1200, and 1800 cycles to reset the initial parameter values. As a result, the SOC estimation performance was improved to within RMSE 5% after 600 cycles. The SOC estimation performance and the SOC estimation errors at 0, 600, 1200, and 2100 cycles are illustrated in Figure 15. Internal resistance estimation showed no significant impact compared to the previous method, as confirmed in Figure 16.

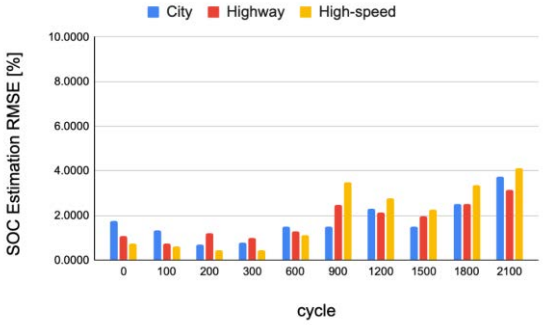
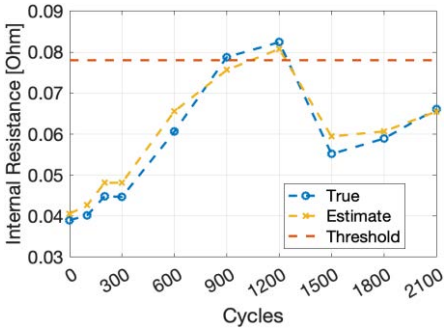
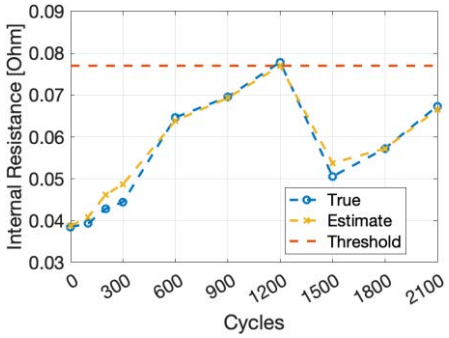


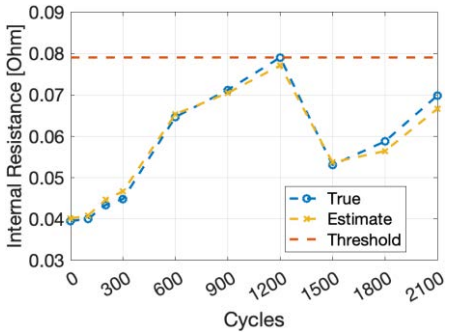
Figure 15. Estimation results of SOC using DEKF, FPIM and initial parameter estimation update



(a)



(b)



(c)

Figure 16. Resistance estimation results of PHEV datasets (a) City, (b) Highway and (c) High-speed

5. CONCLUSION

In this study, a methodology is investigated to estimate SOC and SOH of the battery whose capacity degraded from 0 to 30%. A hybrid approach is proposed that the DEKF and the capacity estimation technique are incorporated to overcome the inherent weak observability of the capacity in the DEKF algorithm. The capacity estimation is implemented by the fixed-point iteration method (FPIM) during the battery charging, and the resulting estimated capacity value is held constant in the process of DEKF during battery discharging. However, we observed a decline in estimation performance beyond a 10% capacity loss, prompting us to propose an initial parameter estimation update to address this issue. As a result, the proposed methodology achieves an accurate and reliable co-estimation of SOC and SOH, even in the battery aging with SOC estimation error lower than 5% and SOH estimation error lower than 3%, even for a battery cell with a capacity fade of 30% for three profiles including hybrid operation modes.

REFERENCES

Campestrini, C., Heil, T., Kosch, S., & Jossen, A. (2016). A comparative study and review of different Kalman filters by applying an enhanced validation method. *Journal of Energy Storage*, *8*, 142–159.

Guo, H., Wang, Z., Li, Y., Wang, D., & Wang, G. (2017). State of charge and parameters estimation for Lithium-ion battery using dual adaptive unscented Kalman filter. *2017 29th Chinese Control And Decision Conference (CCDC)*, 4962–4966.

Guo, R., & Shen, W. (2022). A model fusion method for online state of charge and state of power co-estimation of lithium-ion batteries in electric vehicles. *IEEE Transactions on Vehicular Technology*, *71*(11), 11515–11525.

Hannan, M. A., Lipu, M. S. H., Hussain, A., & Mohamed, A. (2017). A review of lithium-ion battery state of charge estimation and management system in electric vehicle applications: Challenges and recommendations. *Renewable and Sustainable Energy Reviews*, *78*, 834–854.

Hossain, M., Haque, M. E., & Arif, M. T. (2022). Kalman filtering techniques for the online model parameters and state of charge estimation of the Li-ion batteries: A comparative analysis. *Journal of Energy Storage*, *51*, 104174.

Hu, C., Youn, B. D., & Chung, J. (2012). A multiscale framework with extended Kalman filter for lithium-ion battery SOC and capacity estimation. *Applied Energy*, *92*, 694–704.

Hu, X., Yuan, H., Zou, C., Li, Z., & Zhang, L. (2018). Co-estimation of state of charge and state of health for lithium-ion batteries based on fractional-order calculus. *IEEE Transactions on Vehicular Technology*, *67*(11), 10319–10329.

Lee, S., Kim, J., Lee, J., & Cho, B. H. (2008). State-of-charge and capacity estimation of lithium-ion battery using a new open-circuit voltage versus state-of-charge. *Journal of Power Sources*, *185*(2), 1367–1373.

Li, X., Wang, Z., & Zhang, L. (2019). Co-estimation of capacity and state-of-charge for lithium-ion batteries in electric vehicles. *Energy*, *174*, 33–44.

Ma, L., Xu, Y., Zhang, H., Yang, F., Wang, X., & Li, C. (2022). Co-estimation of state of charge and state of health for lithium-ion batteries based on fractional-order model with multi-innovations unscented Kalman filter method. *Journal of Energy Storage*, *52*, 104904.

Mishra, S., Swain, S. C., & Samantaray, R. K. (2021). A Review on Battery Management system and its Application in Electric vehicle. *10th International Conference on Advances in Computing and Communications, ICACC 2021*. https://doi.org/10.1109/ICACC-202152719.2021.9708114

Plett, G. L. (2004). Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 3. State and parameter estimation. *Journal of Power Sources*, *134*(2), 277–292.

Plett, G. L. (2005). Dual and joint EKF for simultaneous SOC and SOH estimation. *Proceedings of the 21st Electric Vehicle Symposium (EVS21), Monaco*, 1–12.

Plett, G. L. (2006). Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 1: Introduction and state estimation. *Journal of Power Sources*, *161*(2), 1356–1368.

Plett, G. L. (2015). *Battery management systems, Volume I: Battery modeling*. Artech House.

Sepasi, S., Ghorbani, R., & Liaw, B. Y. (2014). A novel on-board state-of-charge estimation method for aged Li-ion batteries based on model adaptive extended Kalman filter. *Journal of Power Sources*, *245*, 337–344.

Shrivastava, P., Soon, T. K., Idris, M. Y. I. Bin, & Mekhilef, S. (2019). Overview of model-based online state-of-charge estimation using Kalman filter family for lithium-ion batteries. *Renewable and Sustainable Energy Reviews*, *113*, 109233.

Shrivastava, P., Soon, T. K., Idris, M. Y. I. Bin, Mekhilef, S., & Adnan, S. B. R. S. (2022). Comprehensive co-estimation of lithium-ion battery state of charge, state of energy, state of power, maximum available capacity, and maximum available energy. *Journal of Energy Storage*, *56*, 106049.

Sung, W., & Lee, J. (2018). Improved capacity estimation technique for the battery management systems of electric vehicles using the fixed-point iteration method. *Computers & Chemical Engineering*, *117*, 283–290.

Wang, Z., Feng, G., Zhen, D., Gu, F., & Ball, A. (2021). A review on online state of charge and state of health estimation for lithium-ion batteries in electric vehicles. *Energy Reports*, *7*, 5141–5161.

Wassiliadis, N., Adermann, J., Frericks, A., Pak, M., Reiter, C., Lohmann, B., & Lienkamp, M. (2018). Revisiting the dual extended Kalman filter for battery state-of-charge and state-of-health estimation: A use-case life cycle analysis. *Journal of Energy Storage*, *19*, 73–87.

Wu, J., Jiao, C., Chen, M., Chen, J., & Zhang, Z. (2019). SOC estimation of li-ion battery by adaptive dual kalman filter under typical working conditions. *2019 IEEE 3rd International Electrical and Energy Conference (CIEEC)*, 1561–1567.

Xiong, R., Sun, F., Chen, Z., & He, H. (2014). A data-driven multi-scale extended Kalman filtering based parameter and state estimation approach of lithium-ion polymer battery in electric vehicles. *Applied Energy*, *113*, 463–476.

Xu, Z., Wang, J., Lund, P. D., & Zhang, Y. (2022). Co-estimating the state of charge and health of lithium batteries through combining a minimalist electrochemical model and an equivalent circuit model. *Energy*, *240*, 122815.

Ye, L., Peng, D., Xue, D., Chen, S., & Shi, A. (2023). Co-estimation of lithium-ion battery state-of-charge and state-of-health based on fractional-order model. *Journal of Energy Storage*, *65*, 107225.

Yoo, M. Y., Lee, J. H., Choi, J.-H., Huh, J. S., & Sung, W. (2023). State-of-Charge Estimation of Batteries for Hybrid Urban Air Mobility. *Aerospace*, *10*(6), 550.

Zhang, X., Wang, Y., Yang, D., & Chen, Z. (2016). An on-line estimation of battery pack parameters and state-of-charge using dual filters based on pack model. *Energy*, *115*, 219–229.

Zubi, G., Dufo-López, R., Carvalho, M., & Pasaoglu, G. (2018). The lithium-ion battery: State of the art and future perspectives. *Renewable and Sustainable Energy Reviews*, *89*, 292–308.

Statistical Knowledge Integration into Neural Networks: Novel Neuron Units for Bearing Prognostics

Thomas Pioger[1], Marcia L. Baptista[2]

[1,2] *Delft University of Technology,*
Building 62 Kluyverweg 1,
2629 HS Delft, Netherlands
t.p.pioger@tudelft.nl
m.lbaptista@tudelft.nl

ABSTRACT

Prognostics and Health Management (PHM) is a framework that assesses the health condition of complex engineering assets to ensure proper reliability, availability, and maintenance. PHM can be used to determine how long a machine can function before failure by predicting the Remaining Useful Life (RUL). Neural networks have been used for RUL prediction, but these data-driven models rely solely on data to explicitly integrate knowledge. Recently, authors have proposed physics-informed neural networks (PINNs) to address this limitation. PINNs are neural networks that incorporate expert knowledge and physics in different ways (observational, inductive, and learning bias). Despite their significance, these models tend to be case-dependent and challenging to configure. In this work, we propose statistical neuron units that can be integrated into any neural network. The proposed neuron units extract features from raw data using various statistical functions. Importantly, these modules can be located in different parts of the neural network, and they can be optimized automatically by backpropagating the modules' weights during training. In a study involving bearing degradation behavior, we compare a classical neural network with our modular version. Our proposed RUL estimation model outperformed the baseline, with a reduction of 13% in the root mean square error and a reduction of 7% in the mean absolute error. We also observe an increase of 40% and 21% for the $\alpha - \lambda$ accuracy metric for an α equal to 0.1 and 0.2 respectively. Our code is available publicly on **Github**.

Keywords: Feature extraction, knowledge integration, optimization of parameters, interpretability, accuracy, modularity, neural network

1. INTRODUCTION

Prognostics and Health Management (PHM) is a critical aspect of modern industrial systems, enabling the early detection of faults and the implementation of timely maintenance and repair strategies. One of the key components of PHM is the prediction of the Remaining Useful Life (RUL), which estimates the time until a system or component fails. Accurate RUL prediction is essential for optimizing maintenance schedules, reducing downtime and costs (Ramezani et al., 2019). To predict the RUL, multiple approaches have been developed, which can be classified as physical models, data-driven methods, and hybrid methods (Hasib et al., 2021; Ferreira & Gonçalves, 2022).

Recently, models infused with domain expertise have received much attention, such as physics-informed neural networks (PINNs), a subfield of neural learning that incorporates explicit prior knowledge (Nguyen et al., 2019). This knowledge can come from two sources: scientific knowledge and expert knowledge (Willard et al., 2022; Kang et al., 2021). Scientific knowledge spans a broad spectrum of domains and engineering disciplines, such as empirical equations (J. Wang et al., 2020) or high-resolution bearing dynamic simulations serving as a method for training the model (Sobie et al., 2018). Expert knowledge refers to knowledge obtained through experience that can be used for various purposes during the process of selecting and developing features.

Despite some successful cases of knowledge integration in data-driven models, some limitations persist (Dourado & Viana, 2020; Nascimento & Viana, 2019). Typically, knowledge inclusion is predetermined and fixed, so it cannot be optimized during training. Another challenge is the interpretability of the model, which remains an issue. The lack of explainability power of neural networks makes it difficult to understand how certain models use knowledge in their predictions (Faroughi et al., 2022).

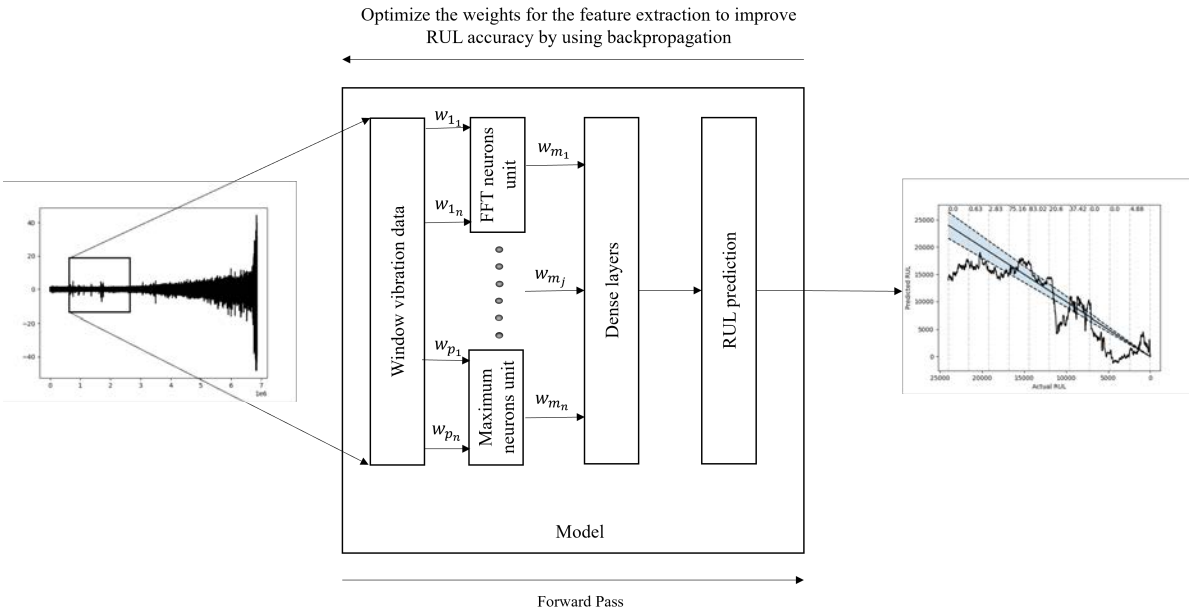


Figure 1. A generic diagram of how one neural network model incorporating the neuron units works. At first, a portion of the vibration signal is fed to the neural network model. This vibration signal is then passed to the different neuron units, which first weight the data received by their own weight and then extract the features needed. The different features extracted are then fed to the dense layers, which are then used to predict the RUL. During training, the neural network model optimizes the weights through backpropagation.

This paper proposes the concept of "knowledge-infused statistics" neuron units for neural networks. These modular units aim to make the structure of neural networks more accurate. Our proposed approach allows the model to train and optimize statistical knowledge during the training stage. Importantly, our proposed neuron units incorporate knowledge that can be fine-tuned and optimized. Since they are modular, these neuron units can be located/ positioned at various locations within the neural network, providing flexibility and adaptability. This novel approach improves neural network performance by optimizing knowledge-infused statistics neuron units.

We investigate the impact of 21 novel neuron units in a bearing case study. Bearings play a crucial role in machinery and mechanical systems, enabling smooth rotation, friction reduction, and support for heavy loads, ensuring operational efficiency and reliability. Our proposed neuron units aim to optimize (improve/enhance) feature extraction during training. The bearing case study is sourced from FEMTO[1] University (Nectoux et al., 2012).

In the implementation of 21 neuron units, we explore the impact of three **novel** neuron configurations: Single Feature Extraction (SFE), Multiple Feature Extraction (MFE), and Weighted Multiple Feature Extraction (WMFE). The SFE type extracts only one feature, while the MFE type extracts multiple features simultaneously. The WMFE layer extracts multiple features and computes/weights the importance of the extracted output features.

The contribution of this paper can be summarized as follows:

- **Modular knowledge-infused statistics Integration:** Introduction of 21 modular and statistical neuron units within a neural network for predicting bearing residual life.
- **Accuracy:** By incorporating these neuron units, we improve the network's ability to extract the features in an optimized way, which, in this case study, led to an improvement in RUL predictions.

In Fig.1, we present a diagram of how our proposed approach works. First, a window of vibration values is given as input to the neural network model. These data are used by the different neuron units to extract features. Each neuron unit has input and output weights (as well as bias) that are used to capture and measure the importance of the features (Fast Fourier transform, skewness factor, maximum amplitude, etc.). The neuron output is given to the dense layers, which then proceed to predict the RUL. The model updates all the weights automatically (by backpropagation) to obtain a better RUL prediction.

The remainder of this paper is organized as follows: Section 2 provides a review of the literature in PHM and the different modeling approaches. Section 3 describes the methodology

[1]FEMTO = Franche-Comté Électronique Mécanique Thermique et Optique

employed in this study and the evaluation metrics. Section 4 presents the case study used in this paper in more detail. Section 5 presents the results obtained and their interpretation. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Prognostics and Health Management (PHM) is a significant area of research that has gained attention in recent years. One of the primary goals of PHM is to predict the Remaining Useful Life (RUL) of engineering systems and components. This is important because it helps to ensure the safety and reliability of these systems, reduce maintenance costs, and optimize maintenance schedules (Guo et al., 2019). To predict the RUL, multiple approaches have been proposed: physics of failure (PoF), data-driven (DD) and hybrid approaches (L. Liao & Köttig, 2014).

The PoF approach in PHM employs mathematical representations that describe the underlying physics of the system under study (Cubillo et al., 2016). To improve the precision of the remaining useful life (RUL) estimate, Q. Wang et al. (2021) proposed a linear mapping technique that directly relates the degradation characteristics of the bearing with its remaining useful life. However, when a linear algorithm is not feasible due to nonlinearities, an alternative approach is needed.

For example, the Extended Kalman Filter (EKF) can be employed to transform the nonlinear problem into a linear one. Singleton et al. (2014) applied an EKF to predict the RUL. However, linearization can lead to unstable filters if the assumption of local linearity is violated, affecting the accuracy and reliability of the estimation process. The unscented Kalman filter (UKF) is also used for RUL prediction (Cui et al., 2019). While prognostics methods based on Kalman filtering approaches can provide precise predictions of the RUL, they typically assume perfect knowledge of the failure system, which is not feasible in most cases.

Another type of algorithm used for physics of failure are particle filters (PF). PF are a type of sequential Monte Carlo method that can effectively handle nonlinear and non-Gaussian degradation processes (Y. Wang et al., 2021). They represent the state of a system using a set of weighted particles, which are updated with new measurements (Elfring et al., 2021). Cai et al. (2020) proposes a similarity-based particle filter method for remaining useful life prediction with improved performance by incorporating historical knowledge and providing probabilistic RUL estimates.

The DD approach relies on the historical data of a system to predict its future state. According to Kefalas et al. (2019), data-driven approaches rely mainly on statistics or machine learning. Statistical models rely on statistical parameters to make predictions (Si et al., 2011). Xiao et al. (2018) proposes a modified duration-dependent hidden semi-Markov model for online machine health prognostics. Jia & Zhang (2019) presented a Bayesian model to reduce model uncertainty for the prediction of RUL.

Artificial Neural Networks (ANN) have been used to estimate RUL (See a review in Ge et al., 2021). Kang et al. (2021) used the Principal Component Analysis (PCA) for data preprocessing and a Multi-Layer Perceptron (MLP) for the prediction of RUL in production lines. Zhao et al. (2019) utilized a recurrent neural network (RNN) to capture temporal dependencies in the degradation process. They first evaluated the trend features to feed their model with the best trends. Zhang et al. (2018) introduced a method to predict RUL of lithium ion batteries using an LSTM.

The dependency on historical run-to-failure (RTF) data is a common issue when implementing DD approaches for RUL prediction. The availability of RTF data is limited, especially for critical components (Hakami, 2024). This limitation poses a significant challenge, as the effectiveness of predictive maintenance, condition-based monitoring, and other DD methods is highly dependent on this historical information. Without comprehensive data on past failures, models may struggle to accurately predict and prevent future breakdowns in crucial equipment.

To overcome the limitations of physical and DD approaches, (hybrid) machine learning models integrating knowledge have been developed (Karniadakis et al., 2021; Dash et al., 2022). This knowledge can be incorporated by transforming the input data, the loss function, or the model. We designate this observational bias, learning bias and inductive bias respectively (Karniadakis et al., 2021). Chao et al. (2022) presents a novel hybrid framework that combines information from physics-based performance models with deep learning algorithms for prognostics. Chen et al. (2022) proposes a model that integrates the knowledge of natural degradation of mechanical components, which is monotonic throughout the life of the bearings and is characterized by temperature signals. Y. Yu et al. (2020) introduced a physics-guided Recurrent Neural Network (RNN) for structural dynamics simulation, where they integrate the underlying physics of structural dynamics into data-enabled machine learning models for the training and prediction of ML models. Xiong et al. (2023) proposed a hybrid framework that combined the controlled physics-informed data generation approach with a deep learning-based prediction model for prognostics.

Physics-Informed Neural Network (PINN), have also been proposed as a way to implement knowledge inside a neural network. X. Liao et al. (2023) introduces a self-attention mechanism into the architecture of the neural network, allowing the mapping of raw data to a hidden state space. Dourado & Viana (2020) presented a PINN modeling approach for the estimation of bias in the prognosis of corrosion fatigue.

The physics-informed layers embed well-understood physical phenomena, and the data-driven layers are used to implement the physical processes that are difficult to model.

Despite the introduction of knowledge within the model, limitations persist (Huang & Agarwal, 2023). Although hybrid models offer the advantage of incorporating knowledge into the learning process, the interpretability of the learned representations and the basis for predictions can still be limited. This lack of transparency can cause problems, especially in critical domains. Neural network interpretability is crucial, as it allows one to explain how and why a neural network produces specific outputs, enhancing trust and understanding. Various methods aim to provide interpretability by visualizing activations, weights, or features and generating textual explanations (Linardatos et al., 2020; Fan et al., 2021).

In the context of Artificial Neural Networks (ANNs), a Modular Neural Network (MNN) can be decomposed into subnetworks based on its connectivity pattern, allowing for a more granular understanding of the network's behavior (Kirsch et al., 2018). Amer & Maul (2019) classified modularization techniques into four main classes (domain, topology, learning, and output), where each class represents the attribute of the neural network manipulated by the technique to achieve modularity. Understanding the modular structure of neural networks can provide insight into their inner workings, making them more interpretable.

This study introduces novel (modular) neuron units that integrate statistical knowledge for neural network training. The concept behind these "neuron units" is their ability to extract essential characteristics (features) from the model during training. Importantly, this feature extraction is automatically optimized by the network. Using these modular layers, we can also visualize the significance of different parts of the input signal (in this case a vibration signal) to predict the RUL.

3. METHODOLOGY

The subsection 3.1 presents our hypothesis for our research framework. Subsection 3.2 describes our approach that we used to test our hypothesis. Subsection 3.3 presents the features that we used to train the models and subsection 3.4 presents how we evaluated the different models.

3.1. Research Hypothesis

We investigated the following research question:

How can we develop (modular) knowledge-infused statistics neuron units for the prediction of RUL?

And with this question, we have the following hypothesis:

A neural network incorporating knowledge-infused statistics neuron units will present an improvement in RUL prediction accuracy.

The use of these "knowledge-infused statistics" neuron units is intended to improve the model at the level of accuracy (and interpretability). By incorporating these novel neurons into the neural network architecture, we aim to facilitate the integration of feature extraction within the model, allowing it to optimize feature selection.

In this paper, we infuse statistical knowledge into the neuron units. We develop 21 neuron units, each incorporating a different and specific statistical feature. With these neuron units we can better understand the contributions of each neuron to the overall prediction performance, enabling more informed decision-making and model refinement. In addition, we can reuse these neuron units in different tasks (Castillo-Bolado et al., 2021). The statistical knowledge that is implemented is generic (max, min, Fourier transform, etc.). We can position the neuron units in different locations within the neural network which can result in multiple model configurations.

We evaluate our hypothesis on the PRONOSTIA bearing data provided by FEMTO (Nectoux et al., 2012). This dataset constitutes a prognostics case study for bearings based on laboratory RTF vibration signals. The PRONOSTIA dataset is explained in more detail in Section 4.

3.2. Modular Approach

In this study, we use a Multi-Layer Perceptron (MLP), a type of neural network widely used in artificial intelligence (Park & Lek, 2016). The connections between neurons are defined by weights, and the output signals are determined by the sum of the inputs to the node, adjusted by a nonlinear transfer function known as the activation function.

To train an MLP, features are extracted from the data and then fed to the model. In this paper, we do feature extraction inside the network, by designing modular neurons units. By organizing the feature extraction process into modular units, we aimed to enhance the MLP's capacity to learn and extract relevant features effectively for RUL prediction. Each modular unit acts as a neuron within the MLP, focusing on capturing specific characteristics present in the input data.

For example, we have developed modular neuron units to extract fundamental features such as peak-to-peak amplitudes, frequency domain features, and vibration characteristics. In addition, we incorporated modular neuron units to extract multiple features simultaneously, allowing a more comprehensive representation of the features.

This modular design facilitates the integration of statistical knowledge into the model architecture. The neuron units are called modular because their architecture allows them to be placed in different parts of the model. As these modules are responsible for feature extraction, they are placed after the input layer, and their output is then fed into the hidden layers for further processing. Fig.2 illustrates how these modular

neuron units are implemented within the MLP.

Each neuron unit has input- and output trainable weights that are optimized during the training process. The input weights equal the input size. In this case, we fed an input with 500 vibration values. We have chosen 500 for computational effectiveness. The neuron unit have 500 trainable weights. We initialize the weights with ones as values.

The neurons multiply the inputs by the weights and extract the features from these weighted inputs. We have three types of neuron unit: single feature extraction (SFE), multiple feature extraction (MFE) and weighted multiple feature extraction (WMFE).

The SFE neuron unit extracts one feature and performs a single extraction operation. For example, to extract features from the frequency domain, the vibration raw dataset is fed to the Fast Fourier transform neuron units and then given to the other neuron units to extract features. As this type of neuron unit performs a single extraction or operation, we call it Single Feature Extraction (SFE).

The MFE neuron unit integrates multiple features inside of it and extracts an array of features. In this case, there are two variants: one in which the features are then fed to the dense layers and one in which the feature arrays are multiplied by a weighted array. This array modifies the values of the features in a way that allows the model the possibility to choose which one was more impactful for RUL prediction. We designate the first version Multiple Feature Extraction (MFE), and the second one is presented as Weighted Multiple Feature Extraction (WMFE).

The first type of module (SFE) is the most simple, as the weights are updated only to extract one feature or perform one operation. In contrast, in the second one (MFE), the weights are updated to extract multiple features efficiently. The third option (WMFE) is the most complex. We created these three modules to extract multiple features classified into fundamental, frequency, and vibration features.

In this work, we focused on the integration of features inside the model; thus, we did not do feature selection. We used the model to optimize feature extraction by itself, where the implementation of trainable weights could help the model do feature selection. However, as they are modular, adding or removing features in the model can be done in a flexible way. These neuron units were not built with the integration of an activation function, as we did not want to force non-linearity on the features extracted. Instead, the dense neurons use the ReLu activation function.

In this paper, we did not study the impact of modular neuron units as the output layer. However, the output layer can be changed to adapt to any necessary prediction. For example, it is possible to implement a modular neuron unit that extracts the minimum as the last layer if the desired prediction is the minimum RUL remaining. The topic of which activation function to use remains a research question for our group.

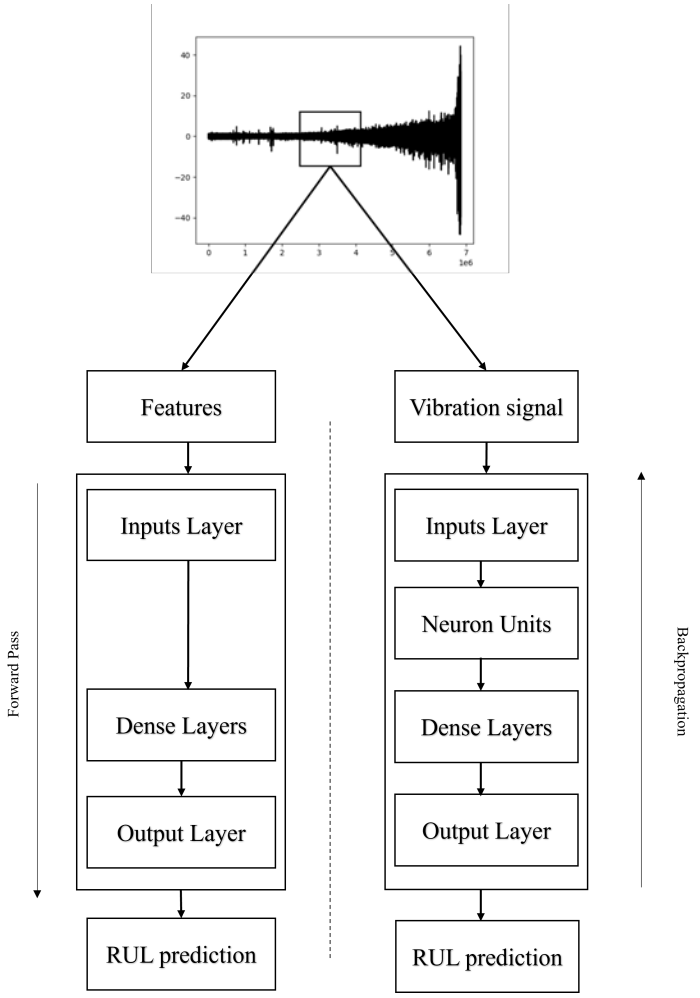
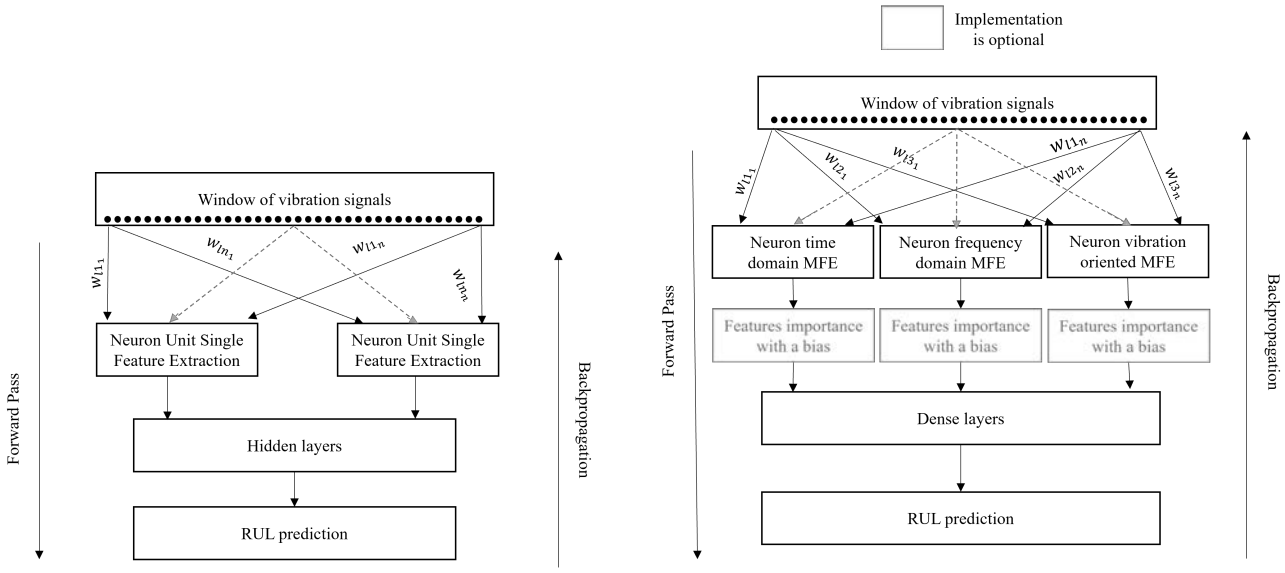


Figure 2. An MLP model without our proposed neuron units on the left and an MLP model with them on the right. The two models have the same number of dense layers with the same number of neurons. The inputs given to the model are different. In the left side model, we feed the model with the features extracted from the vibration signal, while on the right side, we give the model a raw vibration signal.

Table.1 shows the different groups and features used. Fig.3 shows the architecture of the model incorporating SFE, MFE and WMFE neuron units.

3.3. Feature Selection

We utilize multiple features to predict the Remaining Useful Life (RUL). Initially, we implemented classical statistical features, which were extracted from the raw time series data of the vibration signal. These features were termed fundamental because of their general nature. Subsequently, we implemented features in the frequency domain. Initially, we uti-



(a) SFE model

(b) MFE and WMFE model

Figure 3. Figure (a) details how a Single Feature Extraction (SFE) neuron units work inside a MLP. Each SFE has its own weights that they multiply by the inputs received. Figure (b) shows how a Weighted Multiple Feature Extraction (WMFE) and Multiple Feature Extraction (MFE) work. As in the SFE, each neuron unit has its own weights that they multiply on the inputs. However, they extract multiple features in one neuron unit, features that are weighted in a WMFE.

lized the Fast Fourier Transform (FFT) to extract frequency features. Then, we computed the magnitude by taking the absolute value of the FFT. Following the magnitude calculation, we were able to compute the Power Spectral Density (PSD) and the Power ratio of Maximum defective frequency to Mean (PMM) (J. Yu, 2011). From the PSD, we extracted the maximum, sum, mean, and the variance. The final feature type was signal features, encompassing general signal-based statistical metrics applicable to any type of signal, including vibration signals (Khlaief et al., 2019). Table 1 displays all the features used. These features were selected because they are usually used for vibration case studies (Riaz et al., 2017).

3.4. Evaluation Methodology

Since we are dealing with a small dataset, we applied a leave-one-out (LOO) strategy to evaluate the models. In the LOO strategy, we remove one of the bearings from the training set and use it as a test set, leaving us with the remaining bearing vibration signals for the training and validation sets. We employed three widely used metrics in the evaluation: the root mean square error (RMSE), the mean absolute error (MAE) and the $\alpha - \lambda$ metric. These metrics provide valuable insights into the accuracy and precision of the models' predictions. The equations for RMSE and MAE are described in Eqs.1 and 2.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

$$\sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

The third assessment method was the α-λ metric, which is a binary measure determining if a prediction at a given time t_λ falls within the α bounds. This metric measures how well predictions remain within an accuracy cone that narrows over time. We split the total time interval into 10 equal time intervals and calculated the percentage of predictions that fell between the $\alpha = 0.1$ and $\alpha = 0.2$ bounds. We have chosen two strict α to evaluate which models predictions were the most accurate, as a higher percentage of predictions inside the cone indicates a more accurate and reliable RUL prediction model.

4. Case Study

To verify the effectiveness of the proposed methodology, we use the PRONOSTIA bearing dataset. PRONOSTIA is an experimentation platform dedicated to testing and validating bearing fault detection. Fig.4 presents an overview of PRONOSTIA.

The PRONOSTIA dataset was part of the IEEE PHM 2012 Prognostic Challenge. PRONOSTIA comprises three main

Table 1. Features used and their formula

Fundamental		**Frequency**		**Vibration**	
Name	Formula	Name	Formula	Name	Formula
Maximum	$\max(x)$	Maximum	$\max(x)$	Peak to Peak	$\lvert \max(x) - \min(x) \rvert$
Minimum	$\min(x)$	Sum	$\sum_{i=1}^{n} x_i$	Cress Factor	$\frac{\max(x)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}}$
Mean	$\frac{\sum_{i=1}^{n} x_i}{n}$	Mean	$\frac{\sum_{i=1}^{n} x_i}{n}$	Shape Factor	$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}}{\frac{1}{n}\sum_{i=1}^{n} \lvert x_i \rvert}$
Variance	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$	Variance	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$	Impulse Factor	$\frac{\max(x)}{\frac{1}{n}\sum_{i=1}^{n} \lvert x_i \rvert}$
Standard deviation	$\sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i-\overline{x})^2}$	PMM	$\frac{\max(x)}{\frac{\sum_{i=1}^{n} x_i}{n}}$	Clearance Factor	$\frac{\max(x)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \sqrt{\lvert x_i \rvert}^2}}$
Root mean square	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$			Skewness	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3/n}{s^3}$
				Kurtosis	$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^4/n}{s^4}$

parts: a rotating part, a degradation part, and a measurement part:

- The rotating part. The asynchronous motor is the actuator that allows the bearing to rotate through gearing and different couplings. The rotation motion of the motor is transmitted through a gearbox, allowing the motor to reach a speed of 2830 rpm. The human-machine interface of PRONOSTIA allows the operator to change the operating condition.
- The degradation part. A radial force is applied to the test ball bearing, thus reducing the bearing's life duration. This radial load is generated by a force actuator in a pneumatic jack.
- The measurement part. The measurement part acquires the bearing's operation condition and the bearing's degradation. The bearing's degradation is based on two types of sensors: vibration and temperature. The acceleration measures are sampled at 25.6 kHz, and the temperature measures are sampled at 10 Hz.

The dataset consists of three different operating conditions, with a total of seventeen run-to-failure vibration signals given, including six training datasets and eleven testing datasets. The dataset is small, and the life duration of a bearing is relatively large (from 1h to 7h) for the sampling rate. In Fig.5, we present two vibration signals. We did not include all six vibration signals to improve clarity.

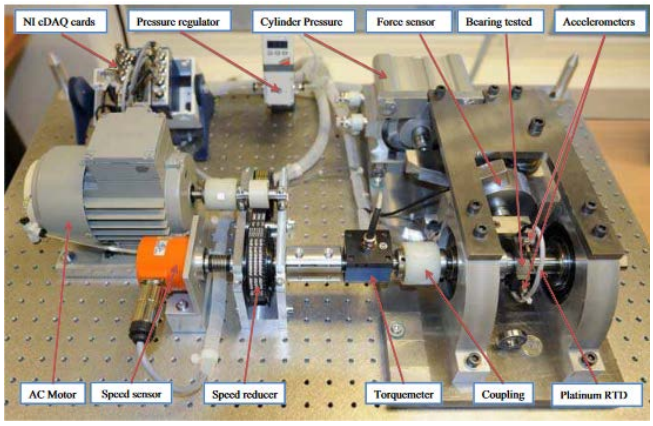
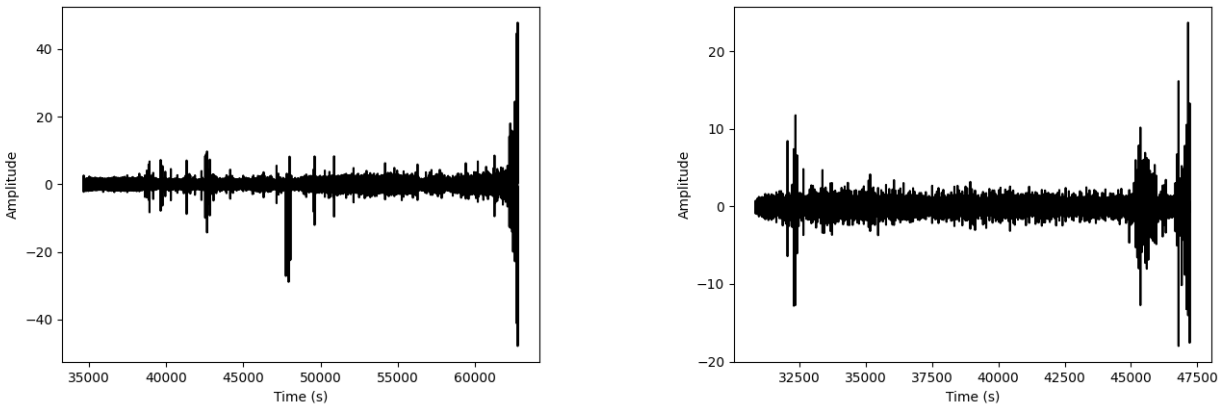


Figure 4. Overview of PRONOSTIA.

5. Results and Discussion

We can confirm our hypothesis: **A neural network incorporating knowledge-infused statistics neuron units will present an improvement in RUL prediction accuracy**. We see in Table.4 and Table.5 on the first bearing that the implementation of MFE neuron units, helped achieve the best $\alpha - \lambda$ score. The MFE model achieved the best RMSE and MAE for the first bearing, according to Table.2. We can see the predictions made by the different models for the first bearing in Fig.6. We can see clearly that the MFE model prediction is the closest one to the true RUL, followed by the baseline model, and then by the SFE and WMFE models.

For the third bearing, despite the baseline model performing better than the other model on the end-of-life, the SFE model



(a) Bearing 1

(b) Bearing 6

Figure 5. Vibration raw signals of the different bearings.

still performs better on the $\alpha - \lambda$ score. With an $\alpha = 0.2$ the SFE model scores a total $\alpha - \lambda$ score of 29 ± 42 while the baseline model scores a total $\alpha - \lambda$ score of 22 ± 21. However, the baseline, has the lowest RMSE and MAE.

If we compare the mean RMSE and MAE values obtained by the different models, we see that the SFE model obtains the lowest values, as seen in Table.3, whereas the baseline is the second best, the WMFE the third one, and the MFE the last one. For the $\alpha - \lambda$ we achieves the best score with the model incorporating the SFE neuron units, while the MFE model is the second best for a small α while the baseline performs better than the MFE model on a higher α.

The results demonstrate that the implementation of knowledge-informed statistics neuron units present an improvement in RUL prediction accuracy, as we have the MFE outperforming the different models in the first bearings and the SFE performing well on the other bearings. These neuron units leverage statistical properties in the model to enhance the RUL prediction.

By adding these knowledge-infused statistical neuron units, we expect to improve interpretability, as we can study the weight evolution during training. The weight evolution can guide us regarding how the model optimizes the feature extraction to predict the RUL. As we know, the model gives a weight to each vibration value given as input. We could then evaluate which part of the signal is more essential for extracting the features needed for an accurate RUL prediction.

6. Conclusion

The objective of this study was to develop a set of novel neuron units for the classical multi-layer perception (MLP). We have evaluated the importance of having these knowledge-informed neuron units inside a neural network aimed at Remaining Useful Life (RUL) prediction. The neurons were infused with statistical knowledge. Concretely, we have implemented 21 neuron units that capture time domain, frequency domain, and time-frequency domain statistical knowledge. Examples of this type of knowledge are the Fourier transform and kurtosis/skeweness.

By using the proposed neuron units, one can create a neural network that incorporates knowledge in an easy and modular way. To test our methodology, we used our network on a bearing case, PRONOSTIA, to predict the RUL. We have demonstrated that these statistical neuron units improve the model prediction compared to a baseline model with classical feature extraction.

Our results showed that the best overall model was the one that incorporated the single feature extraction (SFE) neuron units. This model was able to outperform the baseline on the overall RMSE, MAE and on $\alpha - \lambda$ accuracy. Regarding the $\alpha - \lambda$ accuracy metrics, the SFE model obtained the best overall accuracy. In contrast, the multiple feature extraction (MFE) obtained the second-best score for a strict α (0.2) and the best $\alpha - \lambda$ accuracy metric for the first bearing.

Despite the good performance of MFE on the first bearing, this approach failed to replicate this performance on the remaining bearings. Importantly, the WMFE obtained the best score at the End-of-Life. However, this model did not achieve good accuracy in the previous time intervals.

One potential reason for the underperformance of the models incorporating MFE and WMFE neuron units could be attributed to their architectural design. As they include more features inside of them, their optimization is more challenging. Another reason is that the selection of the features in-

Table 2. RMSE and MAE results for the predictions made on the different bearings by the different models. WMFE stands for Weighted Multiple Feature Extraction, MFE for Multiple Features Extraction, and SFE for Single Features Extraction.

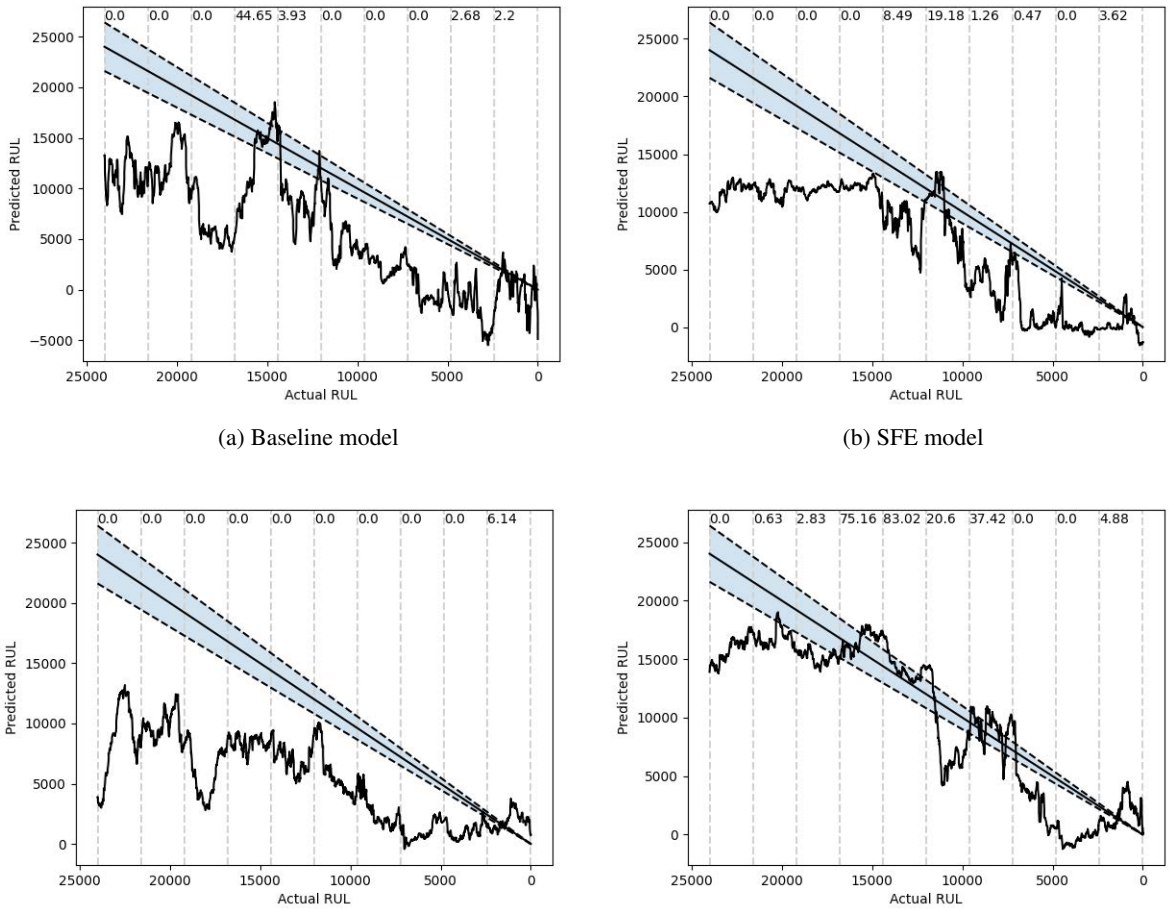
Bearings	**Model with MFE**		**Model with WMFE**		**Baseline**		**Model with SFE**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	**4039**	**3252**	8871	7529	10861	7956	6474	5642
2	11909	9078	5601	4873	**4177**	**3460**	4660	3638
3	7806	6745	8306	7447	3730	2636	**2890**	**2369**
4	11316	9651	7243	5229	**4728**	**3783**	7079	5513
5	16642	15659	9445	9103	6996	5514	**2724**	**2365**
6	14876	13180	9562	8503	**6305**	**5531**	8036	7103

Table 3. Mean and std of the RMSE and MAE for the different models. WMFE stands for Weighted Multiple Feature Extraction, MFE for Multiple Features Extraction, and SFE for Single Features Extraction.

Model with MFE		**Model with WMFE**		**Baseline**		**Model with SFE**	
RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
11098 ± 4211	9594 ± 4047	8155 ± 1378	7114 ± 1568	6133 ± 2405	4813 ± 1757	$\mathbf{5310 \pm 2036}$	$\mathbf{4438 \pm 1776}$

Table 4. Percentage of predictions within the $\alpha = 0.1$ bound. The interval 10 represents the farthest distance from the bearing failure, whereas the interval 1 represents the last interval before failure. WMFE represents the model with Weighted Multiple Feature Extraction units; SFE represents the model with Single Feature Extraction units; B represents the baseline model; and MFE represents the model with Multiple Feature Extraction units.

	Bearings							
	1				**3**			
Interval	WMFE	MFE	B	SFE	WMFE	MFE	B	SFE
10	0.00	0.00	0.00	0.00	0.00	0.00	8.70	**13.04**
9	0.00	**0.63**	0.00	0.00	1.09	1.09	5.98	**66.85**
8	0.00	**2.83**	0.00	0.00	0.00	0.00	5.98	**45.11**
7	0.00	**75.16**	44.65	0.00	0.54	0.00	9.78	**49.46**
6	0.00	**83.02**	3.93	8.49	0.00	0.00	**19.13**	0.00
5	0.00	**20.60**	0.00	19.18	0.00	0.00	**31.15**	0.00
4	0.00	**37.42**	0.00	1.26	0.00	0.00	**5.46**	0.00
3	0.00	0.00	0.00	**0.47**	0.00	0.00	**1.09**	0.00
2	0.00	0.00	**2.68**	0.00	0.00	0.00	0.00	0.00
1	**6.14**	2.20	4.88	3.62	0.00	0.00	0.00	0.00



(a) Baseline model

(b) SFE model

(c) WMFE model

(d) MFE model

Figure 6. Prediction of the different models for the first bearing. The interval represented in light gray is the α bond interval, for $\alpha = 0.1$. Predictions inside this interval are considered correct. The numbers on top represent the correct percentage of prediction inside the α bound for each interval. We compare the baseline prediction with the Single Feature Extraction (SFE), the Weighted Multiple Feature Extraction (WMFE), and the Multiple Feature Extraction (MFE). For clarity of the prediction trend, we are showing here their moving average.

Table 5. Percentage of predictions within the $\alpha = 0.2$ bound. The interval 10 represents the farthest distance from the bearing failure, whereas the interval 1 represents the last interval before failure. WMFE represents the model with Weighted Multiple Feature Extraction units; SFE represents the model with Single Feature Extraction units; B represents the baseline model; and MFE represents the model with Multiple Feature Extraction units.

	Bearings							
	1				**3**			
Interval	WMFE	MFE	B	SFE	WMFE	MFE	B	SFE
10	0.00	0.00	0.00	0.00	2.72	0.00	15.76	**22.28**
9	0.00	**23.90**	0.00	0.00	7.07	4.35	36.41	**97.28**
8	0.00	**34.75**	0.00	0.00	0.00	0.00	20.11	**89.13**
7	0.00	**100.00**	58.18	5.66	1.09	0.00	20.65	**79.89**
6	0.63	**100.00**	9.59	20.13	0.55	0.00	**49.46**	0.00
5	0.00	27.52	1.10	**30.97**	0.00	0.00	**31.15**	0.00
4	0.31	**65.88**	0.00	2.99	0.00	0.00	**16.94**	0.00
3	0.00	0.00	0.00	**0.94**	0.00	0.00	**1.64**	0.00
2	0.16	0.00	**3.46**	0.00	0.00	0.00	0.00	0.00
1	**13.54**	9.61	4.57	6.77	0.00	0.00	**1.09**	0.00

Table 6. Total mean and std of predictions within the α bound for the models.

α	WMFE	MFE	Baseline	SFE
0.1	1.51 ± 2.21	4.50 ± 8.88	3.59 ± 3.00	**5.03± 6.35**
0.2	3.50 ± 5.41	7.76 ± 14.20	7.92 ± 7.13	**9.59 ± 10.35**

cluded in them was made arbitrarily. A way to improve this type of neuron unit is to have a neuron unit composed of all the features instead of splitting them into three different neuron units. This will be researched in future work. The difference between the performance of the MFE and WMFE can be explained by the weights implemented on the feature output array. As the dense layers already have their own weights that are multiplied by the inputs they receive, in this case the feature array, having a weight that does the same operation in the WMFE can be counterproductive.

The proposed models (SFE, MFE and WMFE) were constructed using a typical neuron unit (dense) from TensorFlow, which limits their ability to retain information from prior raw vibration signals and updates the weights solely for a particular time during the bearing's lifespan. As a result, these neuron units might struggle to capture complex patterns in the vibration data. Another limitation of these neuron units is that they need to adhere to the forward and backward propagation mechanisms, which can restrict the complexity of the extracted features.

Another area of optimization can be the placement of neuron units in different locations of the model. In this study, the neuron units were only added at the beginning of the model, after the input layers but before the hidden layers. We can also study the impact of our neuron units at the output layer. For example, because we are predicting the RUL, the minimum neuron unit can be used as the output layer, as the model is attempting to predict the minimum value of the RUL from the values provided as input.

More research is needed to determine whether the implementation of memory-based modular neuron units can achieve better results. Moreover, given that we are dealing with time series data, changing the model architecture could be beneficial for both the baseline and the proposed models. For instance, incorporating long-short term memory (LSTM) layers could improve the model's ability to capture temporal dependencies and patterns in the data. Additional research may impose constraints on the neuron units trainable parameters, forcing the model to extract features in a manner that differs from the existing neuron units. Given their modular nature, we might consider incorporating them not only after the input layer but also in other parts of the model architecture.

Lately, this model was trained offline, and future research can also focus on how to train these neuron units in an online case study where the data will be fed continuously to the model.

Although the primary focus of this study has been on improving prediction accuracy, we recognize the importance of interpretability and aim to leverage knowledge-infused neuron units as a stepping stone towards more transparent and explainable RUL prediction models. Future research efforts will explore techniques to further enhance the interpretability of these models, for example, by the implementation of dif-

ferent neuron units, tracking the weights value during training, or by creating different neurons units that can replace the usual dense layers.

The contribution of this research is the proposal of knowledge-informed neuron units infused with statistical knowledge. These neuron units implement a novel method of extracting statistical features and feeding them to a model, in which the network optimization by backpropagation has a greater impact on the statistical features extracted than if they were directly fed to the model.

ACKNOWLEDGMENT

We would like to thank Dr. Manuel Arias Chao and Kristupas Bajarunas for their help and feedback on this work.

REFERENCES

Amer, M., & Maul, T. (2019). A review of modularization techniques in artificial neural networks. *Artificial Intelligence Review*, *52*, 527–561.

Cai, H., Feng, J., Li, W., Hsu, Y.-M., & Lee, J. (2020). Similarity-based particle filter for remaining useful life prediction with enhanced performance. *Applied Soft Computing*, *94*, 106474.

Castillo-Bolado, D., Guerra-Artal, C., & Hernández-Tejera, M. (2021). Design and independent training of composable and reusable neural modules. *Neural Networks*, *139*, 294–304.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Chen, X., Ma, M., Zhao, Z., Zhai, Z., & Mao, Z. (2022). Physics-informed deep neural network for bearing prognosis with multisensory signals. *Journal of Dynamics, Monitoring and Diagnostics*, 200–207.

Cubillo, A., Perinpanayagam, S., & Esperon-Miguez, M. (2016). A review of physics-based models in prognostics: Application to gears and bearings of rotating machinery. *Advances in Mechanical Engineering*, *8*(8), 1687814016664660.

Cui, L., Wang, X., Xu, Y., Jiang, H., & Zhou, J. (2019). A novel switching unscented kalman filter method for remaining useful life prediction of rolling bearing. *Measurement*, *135*, 678–684.

Dash, T., Chitlangia, S., Ahuja, A., & Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, *12*(1), 1040.

Dourado, A. D., & Viana, F. (2020). Physics-informed neural networks for bias compensation in corrosion-fatigue. In *Aiaa scitech 2020 forum* (p. 1149).

Elfring, J., Torta, E., & Van De Molengraft, R. (2021). Particle filters: A hands-on tutorial. *Sensors*, *21*(2), 438.

Fan, F.-L., Xiong, J., Li, M., & Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *5*(6), 741–760.

Faroughi, S. A., Pawar, N., Fernandes, C., Raissi, M., Das, S., Kalantari, N. K., & Mahjour, S. K. (2022). Physics-guided, physics-informed, and physics-encoded neural networks in scientific computing. *arXiv preprint arXiv:2211.07377*.

Ferreira, C., & Gonçalves, G. (2022). Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems*, *63*, 550–562.

Ge, M.-F., Liu, Y., Jiang, X., & Liu, J. (2021). A review on state of health estimations and remaining useful life prognostics of lithium-ion batteries. *Measurement*, *174*, 109057.

Guo, J., Li, Z., & Li, M. (2019). A review on prognostics methods for engineering systems. *IEEE Transactions on Reliability*, *69*(3), 1110–1129.

Hakami, A. (2024). Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Scientific Reports*, *14*(1), 9645.

Hasib, S. A., Islam, S., Chakrabortty, R. K., Ryan, M. J., Saha, D. K., Ahamed, M. H., ... Badal, F. R. (2021). A comprehensive review of available battery datasets, rul prediction approaches, and advanced battery management. *IEEE Access*, *9*, 86166-86193. doi: 10.1109/ACCESS.2021.3089032

Huang, A. J., & Agarwal, S. (2023). On the limitations of physics-informed deep learning: Illustrations using first order hyperbolic conservation law-based traffic flow models. *IEEE Open Journal of Intelligent Transportation Systems*.

Jia, C., & Zhang, H. (2019). Rul prediction: Reducing statistical model uncertainty via bayesian model aggregation. In *2019 caa symposium on fault detection, supervision and safety for technical processes (safeprocess)* (p. 602-607). doi: 10.1109/SAFEPROCESS45799.2019.9213433

Kang, Z., Catal, C., & Tekinerdogan, B. (2021). Remaining useful life (rul) prediction of equipment in production lines using artificial neural networks. *Sensors*, *21*(3), 932.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., et al. (2019). A review: Prognostics and health manage-

ment in automotive and aerospace. *International Journal of Prognostics and Health Management*, *10*(2).

Khlaief, A., Nguyen, K., Medjaher, K., Picot, A., Maussion, P., Tobon, D., ... Cheron, R. (2019). Feature engineering for ball bearing combined-fault detection and diagnostic. In *2019 ieee 12th international symposium on diagnostics for electrical machines, power electronics and drives (sdemped)* (pp. 384–390).

Kirsch, L., Kunze, J., & Barber, D. (2018). Modular networks: Learning to decompose neural computation. *Advances in neural information processing systems*, *31*.

Liao, L., & Köttig, F. (2014). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, *63*(1), 191-207. doi: 10.1109/TR.2014.2299152

Liao, X., Chen, S., Wen, P., & Zhao, S. (2023). Remaining useful life with self-attention assisted physics-informed neural network. *Advanced Engineering Informatics*, *58*, 102195.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Nascimento, R. G., & Viana, F. A. (2019). Fleet prognosis with physics-informed recurrent neural networks. *arXiv preprint arXiv:1901.05512*.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. In *Ieee international conference on prognostics and health management, phm'12.* (pp. 1–8).

Nguyen, D., Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., & Bäck, T. (2019). A review: Prognostics and health management in automotive and aerospace. *International Journal of Prognostics and Health Management*, *10*(2), 35.

Park, Y.-S., & Lek, S. (2016). Artificial neural networks: Multilayer perceptron for ecological modeling. In *Developments in environmental modelling* (Vol. 28, pp. 123–140). Elsevier.

Ramezani, S., Moini, A., & Riahi, M. (2019). Prognostics and health management in machinery: A review of methodologies for rul prediction and roadmap. *International Journal of Industrial Engineering and Management Science*, *6*(1), 38–61.

Riaz, S., Elahi, H., Javaid, K., & Shahzad, T. (2017). Vibration feature extraction and analysis for fault diagnosis of rotating machinery-a literature survey.. Retrieved from `https://api.semanticscholar.org/CorpusID:40027868`

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, *213*(1), 1–14.

Singleton, R. K., Strangas, E. G., & Aviyente, S. (2014). Extended kalman filtering for remaining-useful-life estimation of bearings. *IEEE Transactions on Industrial Electronics*, *62*(3), 1781–1790.

Sobie, C., Freitas, C., & Nicolai, M. (2018). Simulation-driven machine learning: Bearing fault classification. *Mechanical Systems and Signal Processing*, *99*, 403–419.

Wang, J., Li, Y., Zhao, R., & Gao, R. X. (2020). Physics guided neural network for machining tool wear prediction. *Journal of Manufacturing Systems*, *57*, 298–310.

Wang, Q., Xu, K., Kong, X., & Huai, T. (2021). A linear mapping method for predicting accurately the rul of rolling bearing. *Measurement*, *176*, 109127.

Wang, Y., Peng, Y., & Chow, T. W. S. (2021). Adaptive particle filter-based approach for rul prediction under uncertain varying stresses with application to hdd. *IEEE Transactions on Industrial Informatics*, *17*(9), 6272-6281. doi: 10.1109/TII.2021.3051285

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, *55*(4), 1–37.

Xiao, Q., Fang, Y., Liu, Q., & Zhou, S. (2018). Online machine health prognostics based on modified duration-dependent hidden semi-markov model and high-order particle filtering. *The International Journal of Advanced Manufacturing Technology*, *94*, 1283–1297.

Xiong, J., Fink, O., Zhou, J., & Ma, Y. (2023). Controlled physics-informed data generation for deep learning-based remaining useful life prediction under unseen operation conditions. *Mechanical Systems and Signal Processing*, *197*, 110359.

Yu, J. (2011). Local and nonlocal preserving projection for bearing defect classification and performance assessment. *IEEE Transactions on Industrial Electronics*, *59*(5), 2363–2376.

Yu, Y., Yao, H., & Liu, Y. (2020). Structural dynamics simulation using a novel physics-guided machine learning method. *Engineering Applications of Artificial Intelligence*, *96*, 103947.

Zhang, Y., Xiong, R., He, H., & Pecht, M. G. (2018). Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Vehicular Technology*, *67*(7), 5695-5705. doi: 10.1109/TVT.2018.2805189

Zhao, S., Zhang, Y., Wang, S., Zhou, B., & Cheng, C. (2019). A recurrent neural network approach for remaining useful

life prediction utilizing a novel trend features construction method. *Measurement*, *146*, 279–288.

BIOGRAPHIES

Thomas Pioger (MSc. in Aerospace and Computer Engineering, Institut Polytechnique des Sciences Avancées, September 2021) is currently a Ph.D. student at the Aerospace Faculty of TU Delft since 2022. His main research interests are in the area of Modular Neural Networks for Prognostics and Health Management with a particular focus on the Remaining Useful Life prediction.

Marcia L. Baptista (BS and MSc. in Informatics and Computer Engineering. Instituto Superior Tecnico, Lisbon, Portugal September 2008) is an Assistant Professor at the Aerospace Faculty of TU Delft since 2020. She holds a PhD from the Engineering Design and Advanced Manufacturing (EDAM) program under the umbrella of MIT Portugal. Her research focuses on the development of prognostics techniques for aeronautics equipment. Her research interests include eXplainable Artificial Intelligence (xAI), machine learning, hybrid modeling, maintenance and prognostics.

SurvLoss: A New Survival Loss Function for Neural Networks to Process Censored Data

Mahmoud Rahat[1], and Zahra Kharazian[2]

[1] *Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden*
mahmoud.rahat@hh.se

[2] *Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden*
zahra.kharazian@dsv.su.se

ABSTRACT

This paper presents SurvLoss, a novel asymmetric partial loss and error calculation function for survival analysis and regression, enabling the inclusion of censored samples. An observation in a dataset for which the complete information regarding an event of interest is not available is called censored. Censored samples are ubiquitous in the industry and play a crucial role in Prognostics and Health Management (PHM) by providing a realistic representation of data, improving the accuracy of analyses, and supporting better decision-making in various industries and the healthcare sector. The proposed approach can effectively equip the conventional regression loss functions such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) with the ability to process censored samples. This can impact the field hugely by providing a more accessible usage of neural network models in survival analysis. The proposed survival loss incorporates censored samples by penalizing predictions outside the censoring region and skipping them otherwise. Then, it uses weighted averaging to aggregate the loss from censored samples with the loss from event samples. Unlike many other methods in the field, the proposed model distinguishes itself by avoiding superficial assumptions and exclusively relies on the available information, considering the entirety of the data.

We compared the proposed loss function with its baseline on two publicly available datasets. The first dataset, called C-MAPSS, is from NASA Turbofan Jet Engines simulation, and the second is a recently published real-world dataset from SCANIA trucks. The goal of both datasets is to predict the remaining useful life (RUL) of the machines. The experimental results show that optimization algorithms for training deep neural networks like Adam can effectively utilize the proposed loss function to calculate gradients, update the model's weights, and reduce training and test errors. Moreover, the proposed model outperformed the baseline by taking advantage of the censored samples. The proposed loss function paves the way for the employment of advanced architectures of neural networks with bigger training sizes in survival analysis.

1. INTRODUCTION

This paper deals with the problem of time-to-event prediction. Specifically, the prediction of time until a component fails or, in other words, the component can no longer perform its intended functionality. The literature has three main directions for tackling this problem. The remaining useful life prediction (Revanur, 2020; Altarabichi, 2020; Karlsson, 2023), risk classification (Rahat, Pashami, Nowaczyk, & Kharazian, 2020), and survival analysis (Wang, Li, & Reddy, 2019). While each of these directions has benefits and drawbacks, a shared challenge among all three approaches is dealing with the censored samples. While some methods, like Cox proportional hazards (Cox, 1972), consider them using the partial likelihood function, most methods simply ignore censored samples. Nonetheless, censoring is an inherent aspect of time-to-event prediction, especially in long-term studies. Censored samples refer to data points for which the event of interest (such as death or failure) has not been observed by the end of the study or at the time of analysis. Censoring can happen, for instance, when the exact time of the event is unknown; typically, since the event has not occurred yet, the subject has been lost to follow-up, or the study finishes before we observe the event. With the censored samples, we have incomplete information about the individual, i.e., we don't know when the event happened. Still, we know the event has not occurred during a specific period.

Generally, three types of censoring are recognized in survival analysis (Kleinbaum & Klein, 1996). The right censoring

where the event of interest has not happened for some individuals by the end of the study, but it may occur at an unknown time in the future. In other words, the event time for these individuals is known to be longer than the observed follow-up time. The left censoring happens when we know that the event of interest happened before a specific time, but we don't know precisely when, i.e., the survival time is shorter than the study's start time. Finally, we have interval censoring, where we only observe the event that has happened within a specific time window, but again, we don't know precisely when. Right censoring is considered the most common type; hence, it is regarded as the primary focus of this paper. It is worth mentioning that although we only looked at the right censoring, the proposed loss function can be generalized similarly to all three types.

Censoring is prevalent in the industry for many reasons, such as the long life of components or changes in the conditions of the equipment, like the expiry of a warranty period. In real-world industrial applications, it is common to access much more censored samples than those that fail. Depending on the cases, the ratio of censored to death varies, but this ratio can reach, for instance, 30 censored samples per 1 death or more. Most available methods in the field struggle with consuming censored samples and end up ignoring these samples. This means that they essentially ignore a significant portion of their available data.

An exception to this general rule is methods based on survival analysis designed to take advantage of the censored information. However, survival analysis is much more common in clinical studies than in industry due to its practical limitations, such as the inability to process big data sizes or their shortage in handling temporal information, like sequences of observations from the same individual over time. Additionally, they mostly rely on the Cox proportional assumption, which is known to be naive and not genuine in many real-world cases. On the other hand, there is no way to validate the survival functions produced by these models as we only have access to the time of the event, and the actual degradation curves are unknown. That is why, in many industrial applications, we rely on the median or the mean point of the survival functions. Concordance index (C-index) (Harrell, Califf, Pryor, Lee, & Rosati, 1982) is the typical survival analysis evaluation metric which only considers the order of the events and is known to be biased (Hartman, Kim, He, & Kalbfleisch, 2023; Alabdallah, Ohlsson, Pashami, & Rögnvaldsson, 2024). Most survival datasets are clinical records of patients with a very small number of data points (around 1000) and features (around 10), and there are limitations regarding the proportion of censored data compared to event instances. In most studies, clinical researchers maintain this percentage below 50%.

Additionally, specific constraints are related to applying neural networks within the survival analysis domain. While a handful of methods have been proposed to merge the capabilities of neural networks with survival analysis (Katzman et al., 2018; Kvamme, Borgan, & Scheel, 2019), the prediction accuracy for deep learning methods remains comparable to the classical methods such as Cox and Random Survival Forest (Ishwaran, Kogalur, Blackstone, & Lauer, 2008) in many datasets. It is shown that the performance gain using deep learning or neural network-based approaches is often around 0.02 to 0.03 in concordance index (Chen, 2020). This is primarily due to various underlying assumptions made by methods, such as the constant ratio of risk over time or the small sizes of the standard survival analysis datasets. On the other hand, the neural network field is growing rapidly, and it is crucial to search for new ways to employ their incredible computational power in fields such as time-to-event prediction.

This paper contributes to the mentioned challenge by introducing a new loss function called survival loss, which essentially enables conventional neural networks to process censored samples along with the standard event samples. The idea is to penalize the model in accordance with the information available. For the event samples, the survival loss performs similarly to an ordinary loss by considering the distance between the model predictions and actual values. For the censored samples, the survival loss only penalizes the model if the predicted value falls outside the censoring time interval. As an example for the right censored samples, the model gets penalized only if its prediction is below censoring time (which we already know the event has not happened in that period). On the other hand, if the model's prediction is larger than the censoring time, the proposed survival loss effectively ignores that sample in the loss calculation as there is no evidence of the precise event time. Ultimately, the final loss value will be reported as a weighted average of the error values from censored and event samples. The weighting of the two error values takes place in accordance with the number of considered samples in each part. The following section defines the proposed loss function in more detail.

2. METHODOLOGY

We first visually explain the intuition behind the proposed survival loss and later define it mathematically. Figure 1 describes how the proposed survival loss function calculates the amount of error for an event sample, right-censored sample, left-censored sample, and interval-censored sample. The follow-up region (represented in red) indicates the period where we monitored the individual, and we know that the event of interest has not occurred inside it. The censoring region (represented in green) displays the period where the event has happened or will happen inside it, but we don't know precisely when. Finally, the red-dotted region represents the time in the future after the event has occurred.

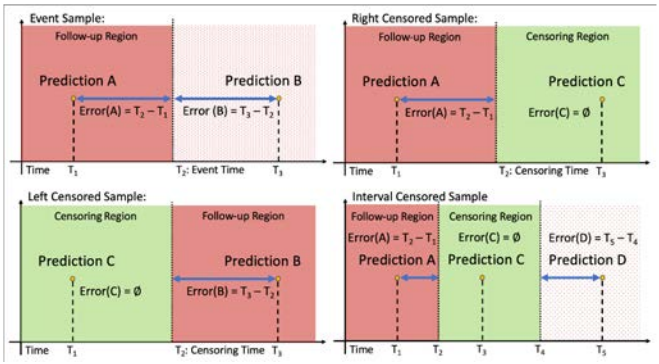


Figure 1. Visual explanation of the proposed loss function.

The loss calculation for an event sample is straightforward. The distance between prediction and actual event time indicates the amount of error. Here, we can use any loss calculation formula like mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), etc. For the sake of simplicity, this paper primarily focuses on the MAE, but any other mentioned error measurement functions can interchangeably be employed.

Unlike the event samples, the proposed loss uses an asymmetric error calculation algorithm for the right censored sample. If the prediction falls within the follow-up region (i.e., the outside of the censoring region), we penalize the prediction according to its distance from the censoring time. Note that this error partially captures the whole error of the prediction as the event happens sometime after censoring time, and the actual error is greater than the proposed partial one. Since calculating the exact error for these samples is impossible, we resort to partial penalization. This partial error from the censored samples can help the optimization algorithm calculate gradients more accurately. Conversely, if the prediction falls within the censoring time, the amount of error is set to null. Note that a null error differs from a zero error since, in the former, we remove the sample from the total number of samples in the batch and, therefore, from the denominator of the averaging function. Finally, a similar logic can also handle the left-censored and interval-censored samples. Again, for the sake of simplicity, we primarily focus on the right censored samples and mean absolute error. Still, the method can be generalized to all censoring types and error calculation functions.

Figure 1 also provides some example prediction points and their associated error values for varying conditions. The errors for prediction A, B, C, and D are respectively $T_2 - T_1$, $T_3 - T_2$, $\emptyset$, and $T_5 - T_4$, where $\emptyset$ means we ignore the sample in the loss calculation.

Assume (X, t, δ) represents a random survival data point where $X \in \mathcal{R}^d$ is a d-dimensional covariate vector and $\delta \in \{0, 1\}$ is an event indicator such that $\delta = 1$, if we observed the event and $\delta = 0$ in case of censoring. Moreover, $t = \min(y, c)$ is the observed time, where $y \in \mathcal{R}^+$ is the actual event time and $c \in \mathcal{R}^+$ is the censoring time.

In the context of PHM and without loss of generality, we define the set of failed samples where the event of failing happened for them (i.e., $\delta_i = 1$) as:

$$(X_i, y_i)_{i=1}^{N_{Failed}} \tag{1}$$

and similarly the set of censored samples (i.e., $\delta_i = 0$) as:

$$\left(\widetilde{X}_j, c_j\right)_{j=1}^{N_{Censored}} \tag{2}$$

which means we divide the samples into two groups of failed (aka event) and censored samples where the total number of samples is $N = N_{Failed} + N_{Censored}$. Then, for a given predictive model f, we define $\hat{y}_i = f(X_i)$ and similarly, $\hat{c}_j = f(\widetilde{X}_j)$ as the output of the predictive model for the failed and censored samples. Note that we have access to the ground truth values for $f(X_i)$, but ground truth values for $f(\widetilde{X}_j)$ are unknown, and the only information we have regarding them is that the actual event time is greater than c_j. This means that we can only penalize the model if the prediction of the model $\hat{c}_j$ is less than c_j; otherwise, we ignore the sample in our loss calculation. Note that ignoring a sample in the loss calculation is different from having an error equal to zero for that sample since by ignoring the sample, we do not consider it in the total number of samples in the denominator of the loss function.

The new survival loss function is defined as a weighted sum of the error for the censored and failed samples:

$$E = \frac{N_{Failed} \times E_{Failed} + N_{Censored'} \times E_{Censored'}}{N_{Failed} + N_{Censored'}} \tag{3}$$

where $Censored'$ represents the set of samples for which the model predicts a survival time less than the censoring time, i.e., $\hat{c}_j < c_j$ and is defined as follows:

$$N_{Censored'} = \left\| \left(\tilde{X}_j, c_j\right) \quad given \quad \hat{c}_j < c_j \right\| \tag{4}$$

this means that $N_{Censored'} \leq N_{Censored}$ since:

$$\{\left(\widetilde{X}_j, c_j\right) \quad given \quad \hat{c}_j < c_j\} \subseteq \{\left(\widetilde{X}_j, c_j\right)\} \tag{5}$$

Then, equations 6 and 7 represent the standard calculation of the mean absolute error for the two groups of $Failed$ and $Censored'$.

$$E_{Failed} = \frac{1}{N_{Failed}} \times \sum_{i=1}^{N_{Failed}} |y_i - \hat{y}_i| \tag{6}$$

$$E_{Censored'} = \frac{1}{N_{Censored'}} \times \sum_{j=1}^{N_{Censored'}} |c_j - \hat{c}_j| \; given \, \hat{c}_j < c_j \tag{7}$$

and if we plug in equations 6 and 7 into equation 3, we get Equation 8 that defines Survival Mean Absolute Error (S-MAE). Equation 8 can easily be modified for Mean Squared Error (MSE) and Root Means Squared Error (RMSE). Note that in the extreme cases of very high censoring ratio, the denominator of the Equation 8 can become zero. To avoid such cases, we recommend reducing the censoring ratio by randomly skipping some of the censored samples or increasing the batch size of the gradient descent algorithm.

3. Experimental Results

We first provide some general information about the experiments and the two datasets used in section 3.1. This is followed by the experimental results and discussion for the first dataset in section 3.2, and that of the second dataset in section 3.3.

3.1. Experimental Setup

For the experiments, we used two public run-to-failure datasets, one on lab-simulated data and the other on real-world data from the field. The first dataset is the well-known NASA Commercial Modular Aero-Propulsion System Simulation, also known as C-MAPSS (Saxena, Goebel, Simon, & Eklund, 2008). It is a widely used benchmark dataset in prognostics and health management (PHM) developed by NASA to support research in aircraft engine health monitoring and prognostics and to estimate the remaining useful life. The dataset consists of simulation measurements from turbofan jet engines with multiple subsets, each corresponding to different operating conditions and engine fault modes. It includes sensor measurements collected from various sensors installed on the engine, along with information about the engine's health and remaining useful life. The C-MAPSS dataset contains temporal information in the form of multiple observations during time from each engine.

The C-MAPSS dataset originally did not contain censored samples as it is a simulated dataset, and the actual failing time for all the engine cases is provided. Since the goal of this paper is to study the effect of having censored samples in the dataset, we used the algorithm introduced in (Rahat, Kharazian, Mashhadi, Rögnvaldsson, & Choudhury, 2023) to transform the dataset into survival settings by defining a specific study period and labeling all the failed samples after the end of the study as censored.

The second dataset is the recently published SCANIA Component X Dataset (Kharazian, Lindgren, Magnússon, Steinert, & Reyna, 2024; Lindgren, Steinert, Andersson Reyna, Kharazian, & Magnússon, 2024). This dataset is collected from an unknown engine component (called component X) of a fleet of trucks. We refer to the second dataset as the Scania dataset. This dataset contains sensor measurements from 21278 censored trucks and 2272 instances of trucks where their component X failed. We define the censoring ratio for a dataset as the percentage of censored samples to the total samples, i.e., the number of censored samples divided by the total number of samples. Therefore, in this dataset, the censoring ratio is 90%, which is way beyond the common censoring ratios in the survival analysis domain. Looking at the literature, it is very rare to see a dataset that contains more than 50% censored samples. Similar to C-MAPSS, this dataset includes temporal measurements from trucks. The only difference compared to C-MAPSS is that in C-MAPSS, the intervals between the observations are the same, but here they vary.

The predictive model used for the experiments is a multilayer Perceptron neural network that contains an input layer followed by five dense layers, each containing 14 neurons, followed by a single neuron as the output, where all layers use the ReLU activation function. The purpose of the network is to predict the remaining useful life of a piece of equipment according to the covariate features received as input. All the networks are trained using the Adam optimization algorithm, and the batch size for all experiments is 32.

There is no need to spend too much time optimizing the neural network's architecture, as both the baseline and the proposed model use the same architecture in terms of fairness of the comparisons. We also tweaked the network's architecture and confirmed that the proposed loss function is not sensitive to the architecture and can perform robustly regardless of its structure. Two models are compared in the following sections. The first model uses the mentioned neural network architecture with S-MAE as the loss function and is referred to as the proposed model. Due to the use of S-MAE, the proposed model can consume censored samples. The second model also uses the mentioned neural network architecture. The only difference is using the standard MAE loss function, which makes the second model unable to render censored samples as there are no ground truth target values associated with them in the dataset. We refer to the second model as the baseline model. The code is implemented in Python using Keras running on a TensorFlow backend. A code implementation of the S-MAE function in TensorFlow is also provided in the appendix section.

$$E_{S-MAE} = \frac{\left(\sum_{i=1}^{N_{Failed}} |y_i - \hat{y}_i|\right) + \left(\sum_{j=1}^{N_{Censored'}} |c_j - \hat{c}_j| \; given \; \hat{c}_j < \; c_j\right)}{N_{Failed} + N_{Censored'}} \tag{8}$$

3.2. SCANIA Dataset

We used the SCANIA dataset for the first experiment. Here, the number of censored to failed samples is enormous, which means most of the components did not fail during the study period. As mentioned before, the ratio of censored vehicles is 90%. In other words, we have around 9 censored vehicles per failed one. Including all censored samples in the experiments is technically impossible, as this will cause the loss to become zero for almost all batches, and consequently, the gradients will become NaN. We randomly down sample censored instances to 500 vehicles. We also include all 2272 instances of failed trucks. The resulting dataset has a censoring ratio of 18%. Additionally, we ignored the temporal information and randomly picked one observation per truck for this experiment. The number of independent features in this dataset is 105, and the goal is to predict the remaining useful life of component X.

The objective of the experiment is to investigate how the inclusion of censored samples impacts a model employing the proposed survival loss function compared to a standard loss function. We conducted two experiments with the model outlined in the experimental setup. In the first experiment, the model is equipped with the ability to incorporate censored samples through the proposed S-MAE loss function. Conversely, the second iteration excludes censored samples from the training data, as the model employs a conventional loss function, making it unable to process partially observed instances. Both models are trained for 10 epochs. The test data used for evaluating both models contains both censored and failed cases, and the S-MAE loss function is used to report the models' performance.

Figure 2 represents the five-fold cross-validation results for the Scania dataset. The two models' average training and test curves across five folds are presented using lines. The shaded confidence bands visualize the respective standard deviations of the test data across five folds. The green curves represent the model's performance using S-MAE, and the red curves represent the conventional MAE loss function. The y-axis shows the value of error using S-MAE. Similar to the standard MAE, the lower values of S-MAE indicate better performances. As you can see in the figure, the model that used the proposed S-MAE consistently outperforms the conventional MAE. Furthermore, the standard deviation of both models decreased during the training epochs.

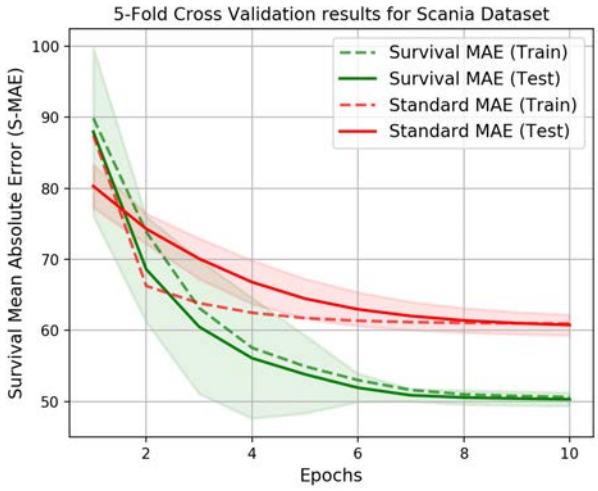


Figure 2. Five-fold cross-validation results for Scania dataset.

3.3. C-MAPSS Dataset

In the second experiment, we performed an assessment similar to the first experiment but with the C-MAPSS dataset. We used the dataset related to the first operational setting (FD001) representing condition one (Sea Level) with 14 covariate features and the remaining useful life of the equipment as the target. The train and test trajectories each contain 100 units, and a varying number of readouts is available in the dataset for each unit. We merged the train and test units from the original dataset to get a dataset with 200 units and reduced the samples by randomly picking 20 readouts from each engine unit. Again, we employed a 5-fold cross-validation approach to evaluate the performance of the proposed method and compare it to our baseline model. Both models are trained for 15 epochs. To simulate censoring, we used the technique described in (Rahat et al., 2023) and set the end-study parameter to 200 in this experiment. The simulation resulted in 1140 event (failed) samples and 646 censored samples with a censoring ratio of 36%.

Figure 3 displays the train (shown with a dashed line) and test (shown with a solid line) learning curves for the baseline and the proposed models. The green curves show the proposed model learning curves, while red is used for the baseline. The shaded area around the test curves represents the standard deviations between the results from the five folds. The standard deviations of the training curves are not visualized to avoid overcrowding and maintain readability. The models are trained until the learning curves flatten after about 15 epochs. As can be seen, the proposed model outperforms the baseline

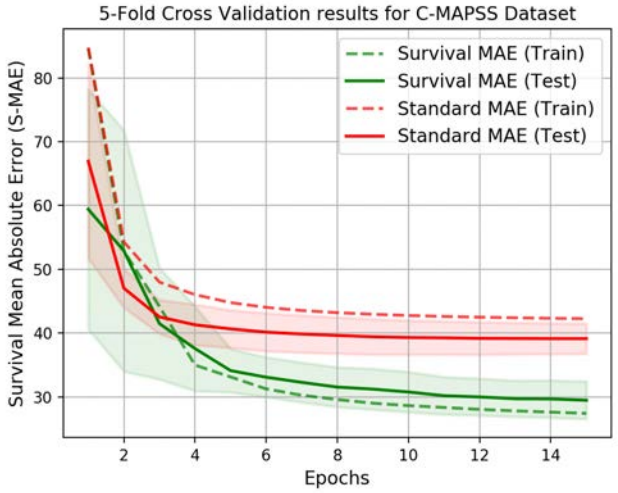


Figure 3. Five-fold cross-validation results for C-MAPSS dataset.

Table 1. The final performance of two models on the test data averaged over 5 folds.

	Scania Dataset	C-MAPSS Dataset
S-MAE	50.32 ± 0.85	29.49 ± 2.64
MAE	60.77 ± 1.31	39.14 ± 2.13

by a significant margin.

Table 1 compared the final performance of two models on the test data averaged over five folds. The proposed loss function significantly outperforms the standard mean absolute error with a margin of 10.45 units in the Scania dataset and with a margin of 9.65 units for the C-MAPSS dataset. There is no need to run statistically significant tests as the standard deviation of the models compared to the net improvement level is little.

4. Conclusion

We presented a novel loss and error calculation method that partially considers censored samples in the context of survival analysis and remaining useful life prediction. The proposed loss function can be used with any standard regression error function and can handle right, left, or interval-censored samples. To assess the algorithm, we tested it using a flat regressor on two public industrial datasets to predict the remaining useful life of engine equipment. The results indicated that the proposed loss function can significantly reduce the model loss on the test data compared to the baseline. The experiments only looked at the mean absolute error function and right censored samples. The application of other regression loss functions with varying censored settings is left for future work. Another suggested future work is to use the proposed loss function on advanced types of neural networks, such as long short-term memory networks (LSTM) or Gated recurrent units (GRU), to incorporate temporal information in survival analysis.

Acknowledgment

The work was supported by research grants from KK-Foundation, Scania CV AB, and the Vinnova Program for Strategic Vehicle Research and Innovation (FFI).

References

Alabdallah, A., Ohlsson, M., Pashami, S., & Rögnvaldsson, T. (2024). The concordance index decomposition: A measure for a deeper understanding of survival prediction models. *Artificial Intelligence in Medicine*, *148*, 102781.

Altarabichi, e. a., Mohammed Ghaith. (2020). Stacking ensembles of heterogenous classifiers for fault detection in evolving environments. In *30th european safety and reliability conference, esrel 2020 and 15th probabilistic safety assessment and management conference, psam15 2020, venice, italy, 1-5 november, 2020* (pp. 1068–1068).

Chen, G. H. (2020). Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine learning for healthcare conference* (pp. 537–565).

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Hartman, N., Kim, S., He, K., & Kalbfleisch, J. D. (2023). Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, *42*(13), 2179–2190.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

Karlsson, e. a., Nellie. (2023). Baseline selection for integrated gradients in predictive maintenance of volvo trucks' turbocharger. In *Vehicular 2023-iaria.*

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, *18*, 1–12.

Kharazian, Z., Lindgren, T., Magnússon, S., Steinert, O., & Reyna, O. A. (2024). Scania component x dataset: A real-world multivariate time series dataset for predictive maintenance. *arXiv preprint arXiv:2401.15199*.

Kleinbaum, D. G., & Klein, M. (1996). *Survival analysis a self-learning text*. Springer.

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-

event prediction with neural networks and cox regression. *Journal of machine learning research*, *20*(129), 1–30.

Lindgren, T., Steinert, O., Andersson Reyna, O., Kharazian, Z., & Magnússon, S. (2024). *SCANIA Component X Dataset: A Real-World Multivariate Time Series Dataset for Predictive Maintenance.* Scania CV AB. Retrieved from https://doi.org/10.58141/1w9m-yz81 doi: 10.58141/1w9m-yz81

Rahat, M., Kharazian, Z., Mashhadi, P. S., Rögnvaldsson, T., & Choudhury, S. (2023). Bridging the gap: A comparative analysis of regressive remaining useful life prediction and survival analysis methods for predictive maintenance. In *Phm society asia-pacific conference* (Vol. 4).

Rahat, M., Pashami, S., Nowaczyk, S., & Kharazian, Z. (2020). Modeling turbocharger failures using markov process for predictive maintenance. In *30th european safety and reliability conference (esrel2020) & 15th probabilistic safety assessment and management conference (psam15), venice, italy, 1-5 november, 2020.*

Revanur, e. a., Vandan. (2020). Embeddings based parallel stacked autoencoder approach for dimensionality reduction and predictive maintenance of vehicles. In *Iot streams for data-driven predictive maintenance and iot, edge, and mobile for embedded machine learning: Second international workshop, iot streams 2020, and first international workshop, item 2020, co-located with ecml/pkdd 2020, ghent, belgium, september 14-18, 2020, revised selected papers 2* (pp. 127–141).

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, *51*(6), 1–36.

APPENDIX

SurvLoss Mean Absolute Error (S-MAE) implementation in Python using TensorFlow. Note that S-MAE receives three values as input, ground truth observed times, predictions, and censoring flag that is either True or False.

```python
import tensorflow as tf

def SurvLoss_MAE(y_true, y_pred, censore_flags):
    y_true_event = y_true[censore_flags]
    y_pred_event = y_pred[censore_flags]
    y_true_censored = y_true[~censore_flags]
    y_pred_censored = y_pred[~censore_flags]
    y_pred_event = tf.squeeze(y_pred_event)
    count1 = y_true_event.shape[0]
    error1 = tf.reduce_mean( tf.abs( y_true_event - y_pred_event ) )
    error1 = tf.where(tf.math.is_nan(error1), tf.zeros_like(error1), error1)
    y_pred_censored = tf.squeeze(y_pred_censored)
    mask = tf.cast(y_true_censored > y_pred_censored, tf.float32)
    count2 = tf.reduce_sum(mask)
    error2 = tf.reduce_mean( tf.abs( tf.multiply( mask, (y_true_censored - y_pred_censored) ) ) )
    error2 = tf.where(tf.math.is_nan(error2), tf.zeros_like(error2), error2)
    survloss_mae = (error1 * count1 + error2 * count2) / (count1 + count2)
    return survloss_mae
```

System-level Probabilistic Remaining Useful Life Prognostics and Predictive Inspection Planning for Wind Turbines

Davide Manna[1,2], Mihaela Mitici[2], Matteo Davide Lorenzo Dalla Vedova [3]

[1] *Faculty of Science, Utrecht University, Utrecht, 3584 CC, the Netherlands*
m.a.mitici@uu.nl

[2] *Department of Mechanical and Aerospace Engineering, Politecnico di Torino, 10129, Turin, Italy*
davide.manna@studenti.polito.it; matteo.dallavedova@polito.it

ABSTRACT

Wind energy plays a crucial role in the energy transition. However, it is often seen as an unreliable source of energy, with many production peaks and lows. Some of the drivers of uncertainty in energy production are the unexpected wind turbine (WT) failures and associated unscheduled maintenance. To support an effective health management and maintenance planning of WTs, we propose an integrated data-driven framework for Remaining Useful Life (RUL) prognostics and inspection planning of WTs. We propose a Long-short term memory (LSTM) neural network with Monte Carlo dropout to estimate the distribution of the RUL of WTs, i.e. we develop probabilistic prognostics. Different from existing studies focused on prognostics for single components, we consider the simultaneous health-monitoring of multiple components of the WTs, thus seeing the turbine as an integrated system. The obtained prognostics are further included into a stochastic planning model which determines optimal moments for inspections. For this, we pose the problem of WT inspections as a renewal reward process. We illustrate our framework for four offshore WTs which are continuously monitored by Supervisory Control and Data Acquisition (SCADA) systems. The results show that LSTMs are able to estimate well the RUL of the WTs, even in the early phase of their usage. We also show that the prognostics are informative for maintenance planning and are conducive to conservative inspections.

1. INTRODUCTION

The current global environmental crisis has prompted the active shift towards renewable energy solutions. For this, as outlined in the European Green Deal, the primary objective set forth by the Global Wind Energy Council is to actively contribute to meeting, by 2030, no less than 20% of the worldwide demand for electricity through the utilization of wind energy. Furthermore, the overarching ambition extends to realizing a fully decarbonized electricity supply by 2050, positioning wind energy at the forefront of renewable sources (Apunda & Nyangoye, 2017).

The focus on wind energy is motivated by the fact that wind is a clean, sustainable and inexhaustible source of energy, it has low operational costs, and that WTs can be installed in various locations, including remote areas where higher wind speeds can result in a higher energy production. Wind energy is, however, perceived as an unreliable source of energy, with many production peaks and lows. Some of the main drivers of uncertainty in energy production are the amount of unexpected failures and associated unscheduled maintenance (Letcher, 2023).

Horizontal Axis Wind Turbines (HAWTs), currently the most promising global wind energy technology (Rezamand et al., 2020), often face accelerated degradation due to their placement in regions with harsh and variable meteorological conditions (Astolfi, Pandit, Terzi, & Lombardi, 2022). Exposed to variable aerodynamic loads and mechanical stress (Tchakoua et al., 2014), WTs necessitate continuous health monitoring and dynamic maintenance planning to achieve reliable operations (Yang, Tavner, Crabtree, Feng, & Qiu, 2014; Tautz-Weinert & Watson, 2017).

In general, a HAWT integrates several essential subsystems, including aerodynamic rotor blades, a central hub for energy transfer, a gear reducer (Tong, 2010) (typically spur, helical (Errichello & Muller, 1994), or planetary (Ragheb & Ragheb, 2010)), an electrical generator for power conversion (Wagner, 2020), a nacelle housing all critical machinery, a yaw system enabling optimal wind alignment and a towering structure (Griffith et al., 2016). The subsystems with the highest fault rates for onshore wind farms are towers, gearboxes, and rotor blades, while for offshore wind farms, the most affected

components are gearboxes, rotor blades, generators, and towers (Rezamand et al., 2020). Generally, the most critical components of WTs are the gearbox, the main bearing, and the blades (Yang, Court, & Jiang, 2013).

To boost the reliability of WTs, recent studies have developed diagnostics and prognostics for components of WTs, focusing particularly on critical components such as gearboxes, main bearings, and blades. Given the increasing availability of condition monitoring measurements, a large fraction of these studies develops data-driven approaches for diagnostics and prognostics using machine learning. For data-driven diagnostics of WTs, frequent approaches are clustering algorithms, Principle Component Analysis (PCA), and Neural Networks. For example, in (Kim et al., 2011), a data-driven, unsupervised clustering algorithm, together with PCA is developed for diagnostics of gearboxes of WTs. Anomalies due to gearbox failures are identified based on measurements related to rotor speed and power production. In (Zaher, McArthur, Infield, & Patel, 2009), a multilayer neural networks is proposed to detect anomalies of the WT gearbox. The main input of the neural network is the temperature of the gearbox. In (Garan, Tidriri, & Kovalenko, 2022), the authors estimate whether the WT will fail or not within the next 60 days using a decision tree. Here, the focus is on optimizing the data preprocessing and feature selected steps of the methodology. A regression mode is proposed in (Orozco, Sheng, & Phillips, 2018) to detect anomalies of gearboxes.

For data-driven Remaining Useful Life prognostics using machine learning, which is also the case of our analysis, existing studies have focused on supervised neural networks. Frequently, vibration and/or Supervisory Control and Data Acquisition measurements are considered as input for these neural networks. Table 1 gives an overview of the main data-driven machine learning approaches, as well as the performance achieved. We note that all these studies focus on specific WT components when developing prognostics. The main components considered for prognostics development are the gearbox and the bearings. Neural networks are a frequently employed approach, which achieves accurate prognostics at various prognostics horizons (e.g., months/days before the actual failure). A recent study (Rajaoarisoa, Randrianandraina, & Sayed-Mouchaweh, 2024) develops a recurrent neural network to estimate the RUL of WTs, following the identification of faults using autoencoders. Completementary to this work, in this paper we propose a Long-short term memory (LSTM) neural network that directly estimates the RUL of the WTs. Here, the health monitoring and generation of RUL prognostics is performed at system level, i.e., the wind turbine is seen as an integrated system. Moreover, existing studies do not consider the development of maintenance planning models for WTs based on prognostics, e.g., predictive inspection planning for wind turbines. To the best of our knowledge, we propose for the first time a maintenance planning model for WTs based on RUL prognostics that are developed using actual measurements and machine learning models.

In this paper, we propose a LSTM neural network for RUL prognostics of WT. As datasets, we consider the recordings of the SCADA systems of the EDP Wind Farm open-source dataset (EDP, 2023). Different from existing studies, our approach involves the simultaneous health monitoring of multiple WT components such as the transformer, the gearbox, the generator, the hydraulic system. Consequently, we define the end-of-life of the WT as the occurrence of the first failure among its components. We use a LSTM neural network to estimate RUL prognostics for the WT seen as an integrated system, i.e., we determine system-level prognostics. By applying Monte Carlo dropout in the testing phase of the LSTM, we quantify the uncertainty associated with these prognostics, i.e., we determine probabilistic RUL prognostics. These prognostics are updated over time, as more measurements become available. The results show that the LSTM network is effective in accurately predicting the RUL of the WTs, even in the early stages of usage. Last, taking into account the obtained probabilistic RUL prognostics, we pose the problem of WT inspections as a renewal reward process and develop a planning model for inspections. The results show that the RUL prognostics support a conservative planning of inspections. This inspection planning is adjusted over time, as prognostics are themselves updated with newly acquired measurements.

The remainder of the paper is as follows. Section 2 introduces the open-source dataset considered for prognostics development. Subsequently, in Section 3.1, the importance of these features is quantified based on their SHAP values, and the most important features are selected for prognostics development. Section 3 proposes a LSTM neural network with Monte Carlo dropout for system-level probabilistic RUL prognostics for WT. Section 4 proposes a stochastic planning model for WT inspections, which integrates the probabilistic RUL prognostics obtained. Numerical results for WT system-level RUL prognostics and WT inspection planning are presented in Section 5. Last, conclusions are provided in Section 6.

2. DATA DESCRIPTION

We consider the Energias de Portugal (EDP) open-source dataset consisting of time-series of sensor measurements recorded for five offshore WT located in the West African Gulf of Guinea in the period 1st January 2017 - 31st December 2017. The information available in the EDP dataset consists of SCADA measurements, meteorological recordings, and the logs of the WT component failures, see also the complete list of measurements (EDP, 2023). The capacity of the each WT is 10MW. The measurements are recorded every 10min. For WT09, the logs recorded concern the Gearbox noise and Pitch position error, which does not indicate a proper fault/damage.

Table 1. Overview of data-driven prognostics for WT components, where ANFIS = Adaptive Neuro-Fuzzy Inference System, (K)ELM = (Kernel) Extreme Learning Machine, NN = Neural Network, SVM= Support Vector Machine; ACC = Accuracy, MA(P)E = Mean Absolute (Percentage) Error, NE = Normalized Error, PRC= Precision, (R)MSE = (Root) Mean Squared Error, SSE= Sum Squared Error.

Reference	Component	Method	Achieved Performance
(Li, Xu, Lei, Cai, & Kong, 2022)	Gearbox	NN	RMSE = 0.0025
(Merainani, Laddada, Bechhoefer, Chikh, & Benazzouz, 2022)	Bearing	NN	RMSE = 0.0025
(Kramti et al., 2021)	Bearing	NN	graphs available
(Elasha, Shanbr, Li, & Mba, 2019)	Gearbox	NN	SSE=661.98
(Pan, Hong, Chen, Singh, & Jia, 2019)	Gearbox	ELM	RMSE=0.91, MAE=0.734, ACC=95.4%
(Carroll et al., 2019)	Gear bearing	NN; SVM	ACC=72%; ACC=60%
(Cao, Qian, & Pei, 2018)	Bearing	SVM	RMSE=16.4, MAPE=42.9%
(Herp, Ramezani, Bach-Andersen, Pedersen, & Nadimi, 2018)	Bearing	NN, GP	0.5 $\leq$PRC$\leq$1
(Kramti, Ali, Saidi, Sayadi, & Bechhoefer, 2018)	Bearing	NN	MSE=0.0023
(Teng, Zhang, Liu, Kusiak, & Ma, 2016)	Bearing	NN	NE= 12.78%
(Chen, Matthews, & Tavner, 2013), (Chen, Matthews, & Tavner, 2015)	Pitch system	ANFIS	ACC$\geq$ 78%, prognostic horizon =21days, ACC$\geq$ 80%, prognostic horizon =14days ACC$\geq$ 86%, prognostic horizon =7days
(Zhao, Liu, Jin, Dang, & Deng, 2021)	Bearing	KELM	4.68% <NE<458.14%

As such, for our analysis, we consider the remaining four WTs (WT01, WT06, WT07, WT11).

Preliminary feature selection

Feature engineering from existing studies on prognostics and diagnostics for WTs (see also Table 1), indicate temperature-related features, production power, the generator and rotor speed rotation as parameters with a high explainability power for failures. In this line, we make a preliminary selection from the available parameters, leading to the following 31 features to be analysed for RUL prognostics: Average Temperature Hydraulic Oil (°C), Max/Min/ Average/STD Generator RPM (rpm), Average Temperature Bearing/ Bearing2 (°C), Average Temperature Generator Phase 1/2/3 (°C), Average Temperature Gearbox Oil (°C), Average Temperature Gearbox Bearing (°C), Average Temperature Nacelle (°C), Max/Min/Average Rotor RPM (RPM), Average Temperature High Volt Transformer Phase1/2/3 (°C), Average Temperature Grid Inverter Phase1 (°C), Average Temperature Controller Top (°C), Average Temperature Controller Hub (°C), Average Temperature Controller VCP (°C), Average Temperature Controller VCP Chokcoil (°C), Average Temperature VCP Cooling Water (°C), Average Temperature VCP Cooling Water (°C), Average Temperature Spinner (°C), Latest Production Total Active Power (Wh), Average Temperature Generator Slip Ring (°C), Average Temperature Grid Rotor Inverter Phase1/2/3 (°C).

3. System-level RUL prognostics for wind turbines

We consider a WT consisting of multiple components. The health of each component is monitored continuously by multiple sensors. We say that the system-level RUL of the WT is the remaining time until the first failure of any one of these components. We are interested in estimating the system-level RUL of the WT based on the sensor measurements recorded. At time step d (dth day), we have available the following measurements for WT $i, i \in \{1, 2, \ldots, n\}$,

$$x_d^i = \{x_{1,d}^i, x_{2,d}^i, \ldots, x_{m,d}^i\}, \quad (1)$$

where m is the total number of considered features and $x_{j,d}^i$ is the measurement corresponding to feature $j, 1 \leq j \leq m$ recorded on day d for WT i.

Then, the actual system-level RUL of WT i at time d is:

$$RUL^a(WT_i) = \min\{\tau(c_1^i) - d, \tau(c_2^i) - d, \ldots, \tau(c_n^i) - d\}, \quad (2)$$

where $\tau(c_j^i), 1 \leq j \leq n$ is the time of failure of component c_j^i of WT i, and n the total number of components of WT i.

We are interested in estimating the system-level RUL of the four WTs in the EDP dataset at various moments (k) in time. Table 2 shows four Cases when each of the WT is the testing set, while the datasets of the remaining three WTs constitute the training and validation sets. The failure of three out of the four WTs is due to a failure of the Hydraulic group. The remaining WT fails due to a failure of the Transformer.

3.1. Feature importance using SHAP values

In Section 2, a total of 31 features has been considered. In this section we quantify the importance of these features for WT system-level RUL estimation using the Shapley additive explanations (SHAP) values (Lundberg & Lee, 2017). SHAP values quantify the impact of a feature on the RUL prognostic. SHAP values are determined as follows:

$$\phi_i = \sum_{S \subseteq Fi} \frac{|S|!(|F| - |S| - 1)!}{|F|!} |f(S \cup \{i\} - f(S))|, \quad (3)$$

Table 2. Overview of data used for testing, training and validation - EDP dataset for WT health monitoring.

	Case 1	**Case 2**	**Case 3**	**Case 4**
Testing	**WT06**	**WT07**	**WT11**	**WT01**
Training	WT01, WT07	WT01, WT06	WT06, WT07	WT06, WT07
Validation	WT11	WT11	WT01	WT11
First fault	Hydraulic Group	Hydraulic Group	Hydraulic Group	Transformer
Actual Lifetime	8 months	6 months	4 months	8 months

with F the set of all features considered for RUL prognostics, $S \subseteq F$ a subset of features obtained from the set F except feature i, and $f(S)$ the expected algorithm output given by the set S of considered features. The SHAP value quantifies the magnitude of the impact, i.e., how much a specific feature value contributes to the accurate estimation of the RUL. A large SHAP value for a given feature indicates a large importance of this feature for the RUL estimation.

For each of the four Cases, we select the 60% most important features of the the total of 31 features, i.e., we select 20 features with the highest SHAP value, see Figures 1-4. The results show that, although the WTs have various components that trigger the failure of the entire system, i.e., either the hydraulic group or the transformer, the average RPM of the generator is the feature with the highest importance for all four WTs. These confirms the findings of existing literature (see also Table 1), that the health condition of the generator is crucial for the overall operation of WTs. Most importantly, these results show that regardless of the failure mode, the WT can be seen as a system and the available measurements can support the development of system-level prognostics.

3.2. Long-short term memory (LSTM) for probabilistic RUL prognostics

Given the long-term dependencies in the measurements, as well as the high nonlinearity of the features, we propose a Long-short term memory (LSTM) with Monte Carlo dropout (Hochreiter & Schmidhuber, 1997) to estimate the distribution of the RUL (probabilistic RUL prognostics) of the WTs in each of the four Cases.

We consider a LSTM consisting of L layers, each consisting of N neurons, and LeakyReLu activation layers (Graves & Graves, 2012). The last layer of the LSTM is a Dense layer, for which a ReLu activation function is assumed. The input gate i_t, the output gate o_t, and the forget gate f_t of the LSTM are defined as follows. The forget gate f_t determines whether to consider or not the previous state c_{t-1}, i.e,

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

where x_t is the current input, h_{t-1} is the previous hidden state, W_f is a trainable weight, b_f is bias. The input gate determines whether to update the state of the LSTM using the current observation, using a sigmoid layer:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i). \tag{5}$$

where x_t is the current input, h_{t-1} is the previous hidden state, W_i is a trainable weight, b_i is the bias. The output gate o_t determines whether the hidden state h_t is passed to the next iteration, i.e.,

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o). \tag{6}$$

where x_t is the current input, h_{t-1} is the previous hidden state, W_o is a trainable weight, b_o is the bias. Table 3 shows the hyperparameters of the considered LSTM.

Table 3. Hyperparameters tuning - LSTM.

Number Layers	4
Neurons Layer 1	128
Neurons Layer 2	64
Neurons Layer 3	64
Neurons Layer 4	64
Dropout rate	0.5
Epochs	40
Batch size	32
Window length	3

Monte Carlo dropout for probabilistic RUL prognostics

Commonly, Monte Carlo dropout is applied in the training phase of the neural networks to avoid overfitting. To obtain the distribution of the RUL, i.e., to obtain probabilistic RUL prognostics, we also apply Monte Carlo dropout in the testing phase of the LSTM. In this line, (Gal & Ghahramani, 2016) shows that such a neural network with Monte Carlo dropout approximates a Bayesian neural network representing a deep Gaussian process.

Let X be the samples in the training set of the LSTM, and let Y be the corresponding RUL values. In a Bayesian neural network, we aim to estimate the posterior distribution $p(y|x, X, Y)$:

$$p(y|x, X, Y) = \int p(y|x, \omega)p(\omega|X, Y)d\omega \tag{7}$$

with ω the weights of the neural network, $p(y|x, \omega)$ the probability that the RUL is y, given the test sample x and the weights ω, and $p(\omega|X, Y)$ the posterior distribution of the weights, given the training samples X and Y.

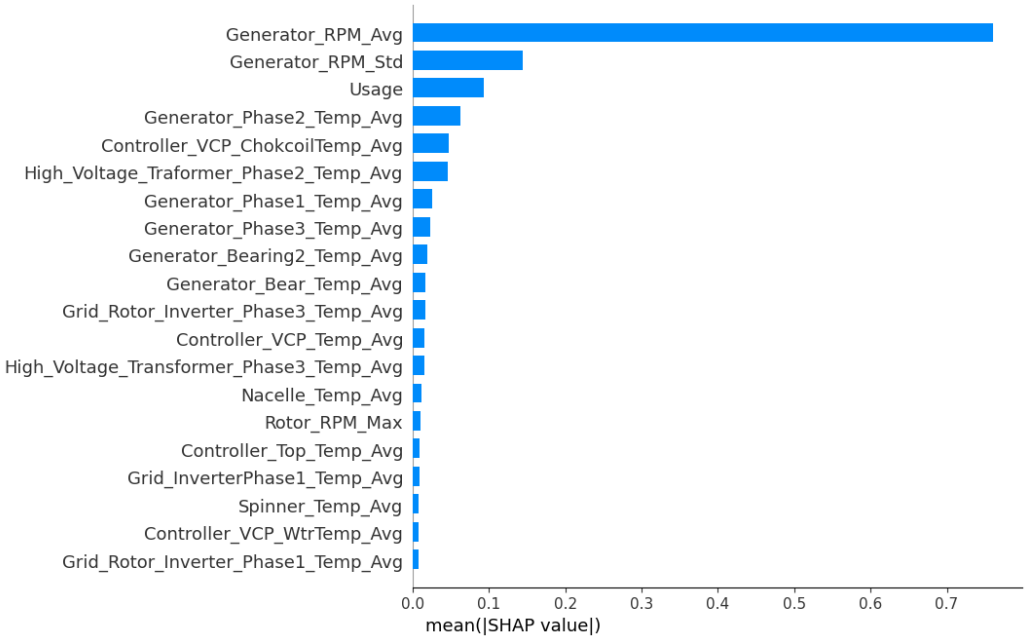


Figure 1. Case 1: WT06 - SHAP values of features.

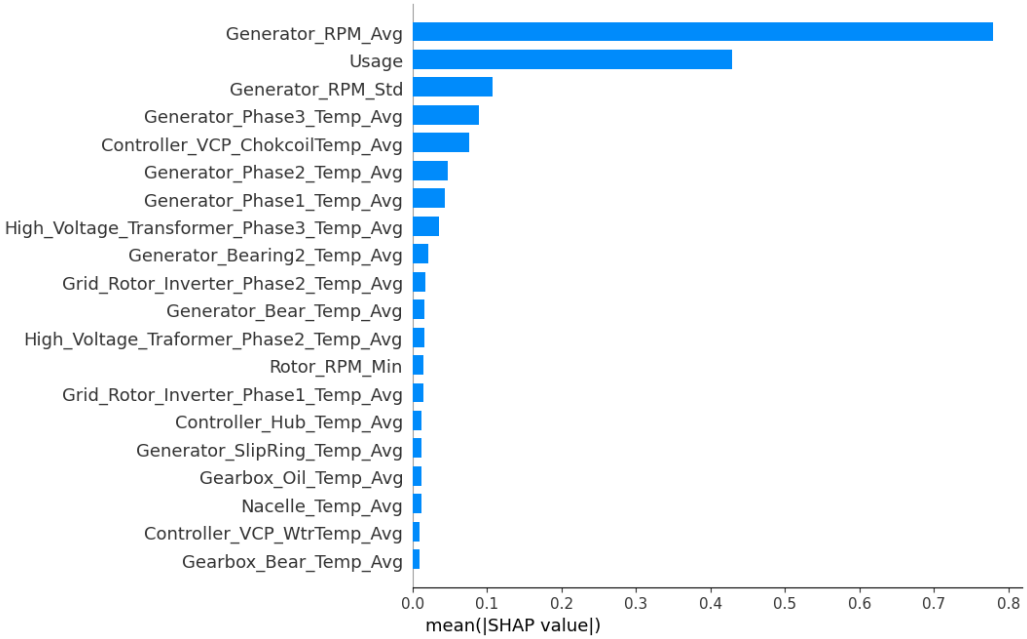


Figure 2. Case 2: WT07 - SHAP values of features.

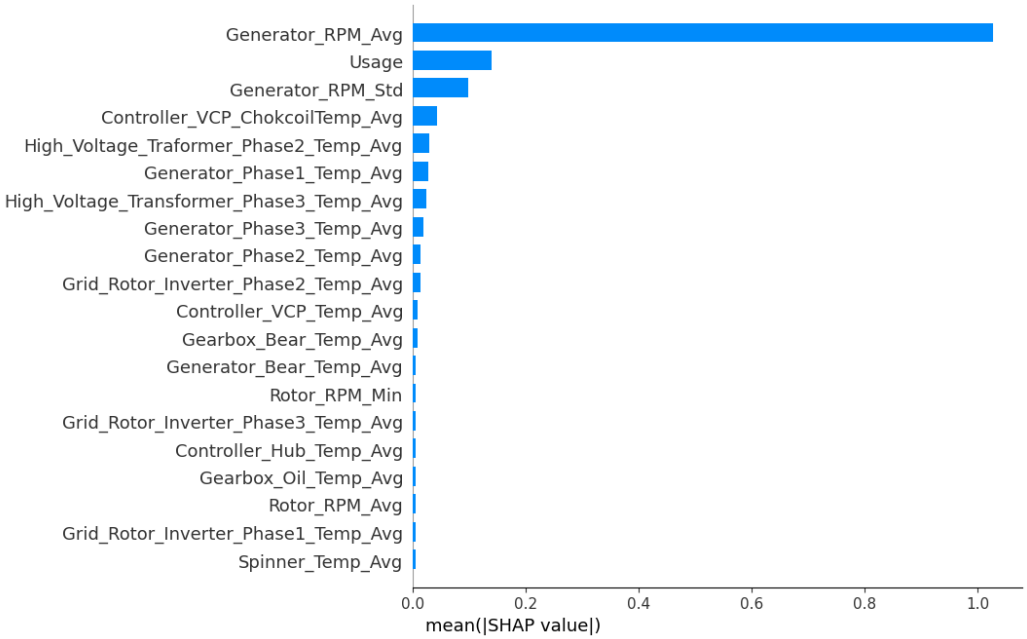


Figure 3. Case 3: WT11 - SHAP values of features.

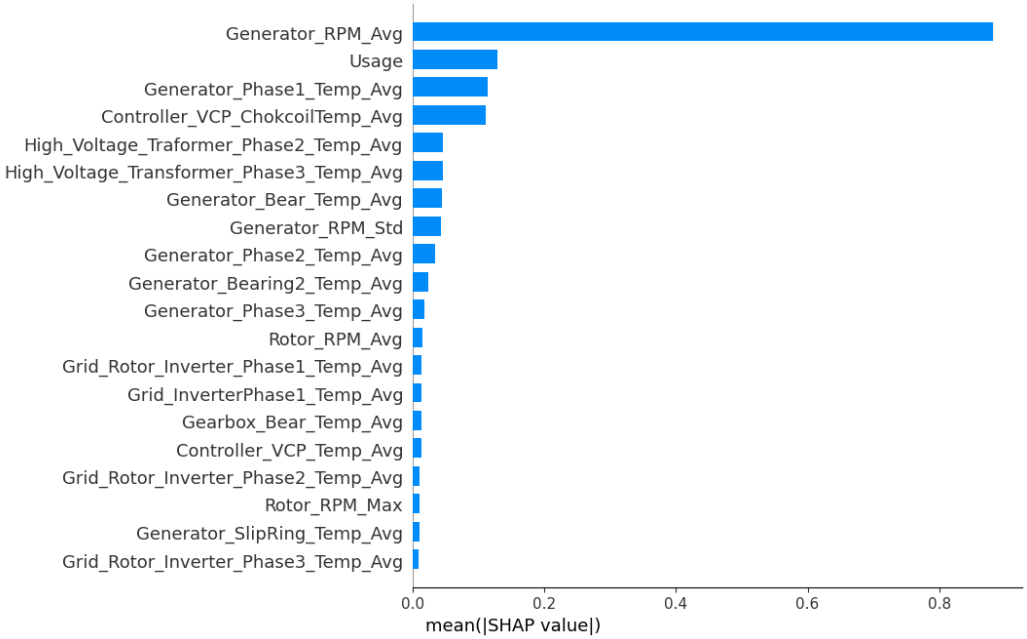


Figure 4. Case 4: WT01 - SHAP values of features.

It is computationally expensive to analyze the posterior distribution $p(\omega|X,Y)$ exactly (Gal & Ghahramani, 2016). As such, we approximate $p(\omega|X,Y)$ with a distribution $q(\omega)^*$ that minimizes Kullback–Leibler divergence KL with the true posterior distribution $p(\omega|X,Y)$, i.e. (Blei, Kucukelbir, & McAuliffe, 2017):

$$q^*(\omega) = argmin_{q(\omega)}\{KL(q(\omega|p(\omega|X,Y)))\}. \quad (8)$$

Using $q(\omega)^*$, we approximate the posterior distribution of the RUL of a test sample by:

$$q(y|x) = \int p(y|x,\omega)q^*(\omega)d\omega \quad (9)$$

where $q(y|x)$ is the approximation of $p(y|x,X,Y)$.

Lastly, we approximate the expected value $\hat{y}$ of the RUL of a test sample by:

$$\hat{y} = E_{q(y|x)}(y) = \frac{1}{M}\sum_{1}^{M}\hat{y_j}(x,\omega^j) \quad (10)$$

where M is the number of forward passes through the neural network, ω^j are the weights of the neural network belonging to the j-th forward pass (i.e., where some neurons are dropped out), and $\hat{y_j}(x,\omega^j)$ is the resulting RUL prediction from the j-th forward pass through the neural network. For the distribution of the RUL, we give each individual RUL prediction $\hat{y_j}(x,\omega^j)$ a probability $\frac{1}{M}$.

Performance metrics for RUL prognostics

To evaluate the ability of the LSTM model to predict the RUL, we consider the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the Continuous Ranked Probability Score (CRPS), which are defined as follows.

$$MAE = \sum_{i=1}^{n}\frac{|RUL_i^a - R\bar{U}L_i^p|}{n}, \quad (11)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(RUL_i^a - R\bar{U}L_i^p)^2}{n}}, \quad (12)$$

with n the number of days over which predictions are made, and $R\bar{U}L_i^p$ the mean predicted RUL at day $i, 1 \leq i \leq n$.

Since we estimate the distribution of the RUL, to be able to quantify the fitness of these distributions relative to the actual RUL (a point value), we consider the Continuous Ranked Probability Score (CRPS) and the Weighted CRPS ($CRPS^W$). Here, CRPS evaluates whether the estimated RUL distribution is centered around the actual RUL of the WT and whether the variance of this distribution is low (a high sharpness of the RUL prognostic) (Mitici, de Pater, Barros, & Zeng, 2023). The Weighted CRPS applies a (larger) penalty β when overestimating the RUL then when underestimating the RUL. This is of particular importance when planning the inspections of the WTs - planning too late inspections (after the actual failure of the wind turbine) does not make effective use of the prognostics to timely identify and act upon anticipated failures of the WTs.

CRPS is defined as follows (Gneiting & Katzfuss, 2014),

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}CRPS_i, \quad (13)$$

$$CRPS_i = \int_{-\infty}^{\infty}(F_{\hat{y}_i}(x) - I\{y_i \leq x\})^2 dx \quad (14)$$

with $I\{y_i \leq x\} = \begin{cases} 1 & \text{if } y_i \leq x \\ 0 & \text{if } y_i > x. \end{cases}$

The weighted CRPS ($CRPS^W$) is defined as follows (Gneiting & Katzfuss, 2014):

$$CRPS^W = \frac{1}{N}\sum_{i=1}^{N}CRPS_i^W, \quad (15)$$

$$CRPS_i^W = (2-\beta)\int_{-\infty}^{y_i}(F_{\widehat{y_i}}(x))^2 dx + \beta\int_{y_i}^{\infty}(F_{\widehat{y_i}}(x) - 1)^2 dx, 0 \leq \beta \leq 2. \quad (16)$$

4. Inspection planning of wind turbines using probabilistic RUL prognostics

In this Section we pose the problem of WT inspections as a renewal reward process (Tijms, 2003), which integrates the probabilistic RUL prognostics developed in Section 3.2. We aim to determine optimal times for WT inspections.

We consider a renewal reward process $\{N_t\}$ where the process regenerates when a wind turbine is inspected, i.e., our knowledge about the actual health condition of the wind turbine is reset upon an inspection. At day k during the life of the WT, we are interested in determining an optimal time $k + t_k^*$ to inspect the WT. At day k, using the measurements recorded up to day k and a LSTM with Monte Carlo dropout (see Section 3.2), we estimate the probability that the RUL of the WT is i days, $i \geq 0$. Let $\phi_k(i)$ denote the probability that the WT, after being used for k days, has a RUL of exactly i days. To determine an optimal time to inspect the WT, we consider the expected cost per unit of time:

$$\frac{\text{[Expected cost over the current inspection cycle]}}{\text{[Expected current inspection cycle]}}. \quad (17)$$

At day k, we are interested in finding an optimal time for

inspection t_k^* such that:

$$t_k^* := argmin_{t_k>0} \frac{\mathbb{E}[C(k,t_k)]}{\mathbb{E}[L(k,t_k)]}, \quad (18)$$

with $C(k,t_k)$ the cost of inspecting the WT at day $k+t_k$, given that this WT has already been used for k days, and $L(k,t_k)$ is the length of the inspection cycle of the WT.

If the WT is scheduled for inspection at day $k+t_k$, then a cost c_r is incurred. If, however, the WT fails at some day $j, k < j < k+t_k$ before an inspection is planned, then a failure cost c_f is incurred (corrective maintenance).

With this, the expected cost over the current inspection cycle of the WT is:

$$E[C(k,t_k)] = c_f \sum_{i=0}^{t_k-1} \phi_k(i) + c_r(1 - \sum_{i=0}^{t_k-1} \phi_k(i)). \quad (19)$$

Also, the expected current inspection cycle is:

$$E[L(k,t_k)] = k + \sum_{i=0}^{t_k-1} i\phi_k(i) + t_k(1 - \sum_{i=0}^{t_k-1} \phi_k(i)). \quad (20)$$

Eq. (18) is solved using a numerical grid search. The estimate $\phi_k(i)$ after every day k is obtained using a LSTM and the methodology in Section 3.2.

5. Numerical results

In this Section we illustrate the results obtained for the probabilistic RUL prognostics and inspection planning for the four WTs for which measurements are available at (EDP, 2023).

5.1. Probabilistic RUL prognostics for wind turbines

Table 4 shows the performance of the system-level RUL prognostics for the WTs in the four Cases considered.

Table 4. Performance - RUL prognostics using LSTM.

	MAE	$RMSE$	$CRPS$	$CRPS^W$ $\beta = 1.9$
Case 1: WT06	12.72	15.52	9.98	2.51
Case 2: WT07	11.30	13.65	7.86	9.16
Case 3: WT11	9.40	11.80	6.93	6.88
Case 4: WT01	19.35	22.42	14.68	3.11

The results show that the lowest MAE and $RMSE$ are obtained for WT11, while the highest MAE and $RMSE$ are obtained for WT01. However, when considering the prognostics as input for inspection planning, we are interested in not missing the failures. This may, however, occur when we overestimate the RUL and, based on these overestimates, we plan late inspections. The Weighted CRPS captures the tendency of the prognostics to overestimate the RUL. We consider a large penalty for RUL overestimation ($\beta = 1.9$), given our ultimate goal of planning inspections for WTs based on prognostics. In this line, we are interested in planning inspection timely, to anticipate the actual failures of the turbines rather than missing these failures.

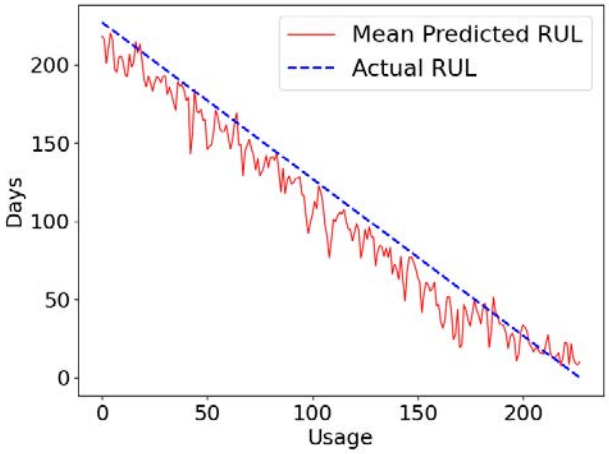


Figure 5. Case 1 - RUL estimation, WT06.

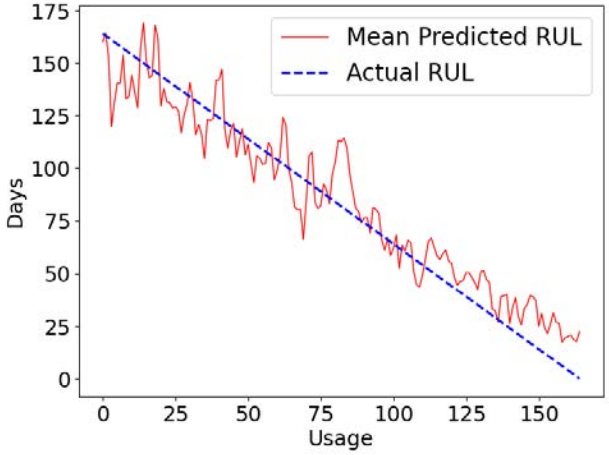


Figure 6. Case 2 - RUL estimation, WT07.

The results show that WT07 has the highest $CRPS^W = 9.16$, despite having a relatively low MAE and $RMSE$. This indicates that the RUL is predominantly overestimated and a conservative inspection planning should be considered, despite the low MAE and $RMSE$. The results also show that WT06 has the lowest $CRPS^W = 2.51$, despite having a relatively high MAE and $RMSE$ among all four turbines. This indicates that the prognostics have the least tendency to overestimate the RUL. These make the prognostics suitable for inspection planning, despite their high MAE and $RMSE$. Overall, the results show that considering MAE and $RMSE$ alone when aiming to use prognostics for maintenance planning is not sufficient. Additional metrics such as $CRPS^W$, through their ability to evaluate whether the

RUL is over/under-estimated, are particularly informative of the suitability of the prognostics for maintenance planning.

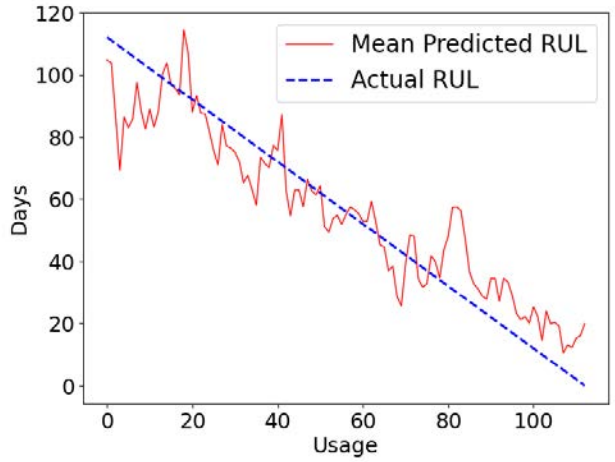


Figure 7. Case 3 - RUL estimation, WT11.

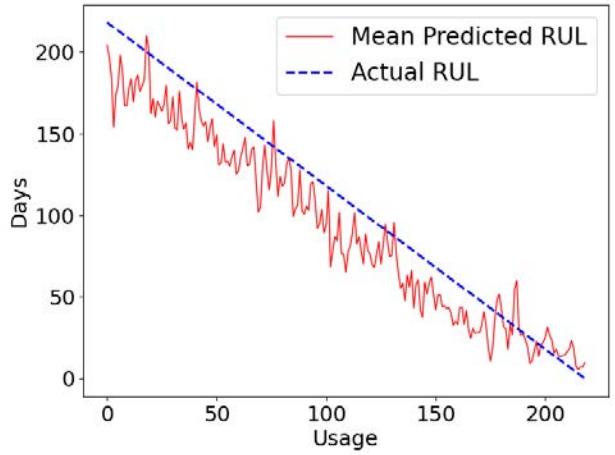


Figure 8. Case 4 - RUL estimation, WT01.

The RUL prognostics obtained over time are shown in Figures 5 - 8. The RUL of WT06 and WT01 are predominantly underestimated. The RUL of WT07 and WT11 are predominantly overestimated.

Figures 9 - 11 show the distribution of the RUL for WT06 (Case 1) at $\{202, 102, 2\}$ days before the actual failure of the WT. The results show that the sharpness of the estimated distribution increases closer to the time of failure of the WT.

5.2. Inspection planning for wind turbines using probabilistic RUL prognostics

For inspection planning, we consider $c_f = 100.000$ and $c_r = 100$. Every day k (or equivalently after k days of usage), based on the measurements collected up to this day, we develop RUL prognostics, i.e., the prognostics are updated ev-

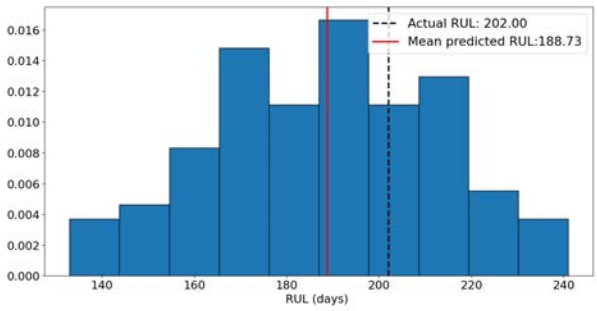


Figure 9. Estimated distribution of RUL, 202 days before the actual failure of WT06.

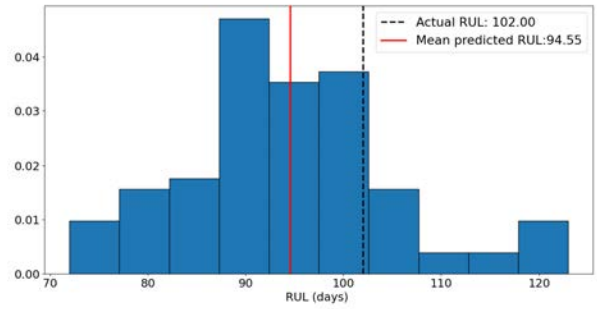


Figure 10. Estimated distribution of RUL, 102 days before the actual failure of WT06.

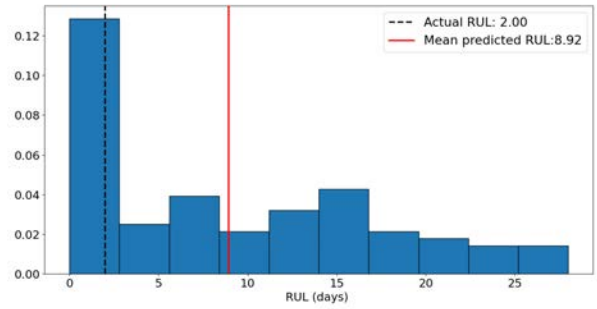


Figure 11. Estimated distribution of RUL, 2 days before the actual failure of WT06.

ery day. Based on these prognostics, every day k we determine an optimal time t_k^* to plan a WT inspection.

Figures 12 -15 show the results for the optimal inspection times of the four WTs relative to the actual RUL and the mean estimated RUL. For Case 1 - WT06, although the MAE and $RMSE$ are relatively high, the fact that $CRPS^W$ is low, i.e., the overestimation of the RUL is low, is reflected in the inspection planning - timely planning that does not miss the failure of the WT. In fact, in the last phase of the monitoring of this WT, it is consistently indicated that an optimal action

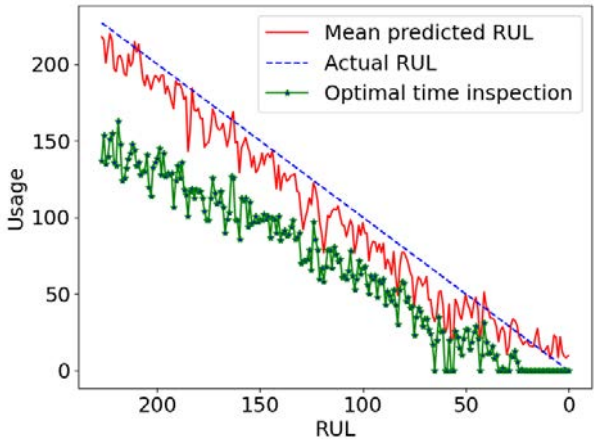


Figure 12. Case 1: WT06, Optimal time inspection.

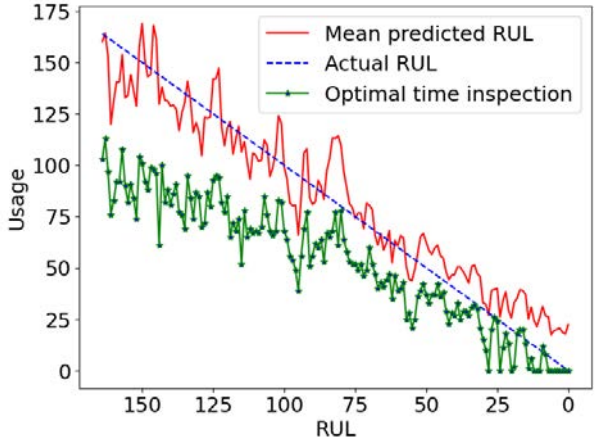


Figure 13. Case 2: WT07, Optimal time inspection.

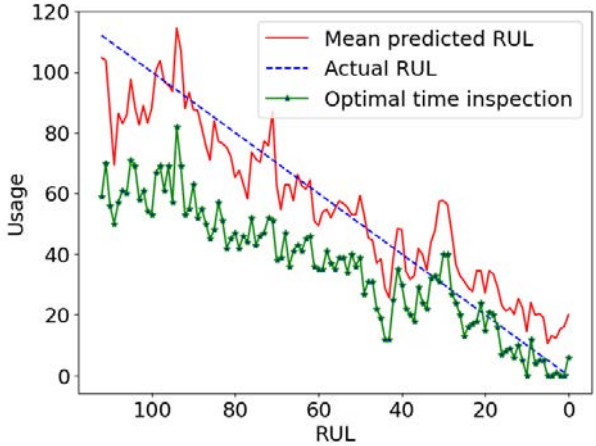


Figure 14. Case 3: WT11, Optimal time inspection.

is to plan an inspection immediately.

For Case 2 - WT07, the $CRPS^W$ is the highest, i.e., the prognostics have the tendency to overestimate the RUL, which is expected to delay the planning of the inspections leading to a potential miss of the failure. This is reflected in the inspection planning, particularly in the last phase of the monitoring of the WT, see also Figure 13. For Case 3 - WT11, the $CRPS^W$ is high, i.e. the prognostics have a tendency to overestimate the RUL. As a result, delayed inspections are planned in the last phase of the monitoring of the WT. For Case 4 - WT01, despite the lowest achieved MAE and $RMSE$, a moderate $CRPS^W$ is reflected in the inspection planning - timely inspection planning, particularly in the last phase of the WT monitoring, when an immediate inspection is consistently indicated as an optimal action (see also Figure 15). Overall, for all four cases, the planning of the inspections is conservative, where timely inspections are indicated as being optimal actions.

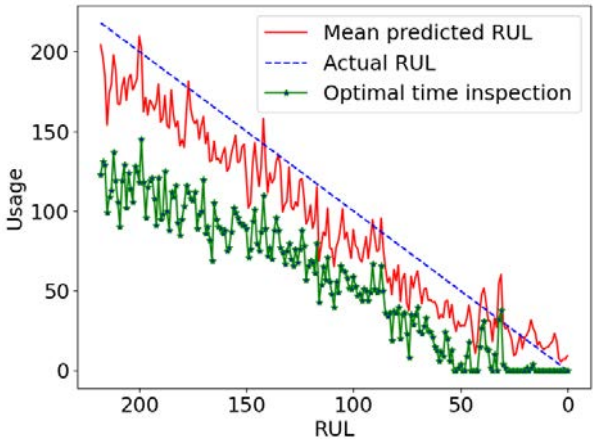


Figure 15. Case 4: WT01, Optimal time inspection.

Table 5. Optimal time for WT inspection.

	RUL^a	k	RUL^p	t_k^*
Case 1: WT06				
	200	27	193.82	139
	100	127	96.62	67
	50	177	49.96	27
	25	202	33.18	6
Case 2: WT07				
	150	14	135.44	101
	100	64	94.11	68
	50	114	38.61	37
	25	139	26.5	24
Case 3: WT11				
	100	12	86.72	53
	75	37	70.28	43
	50	62	61.64	39
	25	87	29.64	13
Case 4: WT01				
	200	18	192.69	118
	100	118	74.69	59
	50	168	28	4
	25	193	8.46	0

Table 5 shows in detail several moments throughout the monitoring of the WTs when inspections are planned (t_k^*), relative to the actual RUL (RUL^a), the usage of the WT (k), and the mean estimated RUL ($R\bar{U}L^p$).

6. Conclusions

This paper proposes a machine learning approach for system-level probabilistic RUL prognostics for WTs. In contrast with existing studies, which develop component-based prognostics, we see the WT as an integrated system and develop system-level RUL prognostics. These prognostics are further employed to determine optimal moments for inspections of the WTs, in anticipation of failures. To the best of our knowledge, this is the first study that proposes a maintenance planning model for WT based on data-driven prognostics. A LSTM with Monte Carlo dropout is developed to estimate the distribution of the RUL of the WTs, i.e., we develop probabilistic RUL prognostics. By using dropout in the test phase of the LSTM, the uncertainty associated with the RUL prognostics is quantified. To plan inspections for the WTs, a renewal reward process is proposed, which integrates these probabilistic RUL prognostics.

We illustrate our approach for four offshore wind turbines located in the West African Gulf of Guinea, and which have been monitored in the period 1st January - 31st December 2017. The results show that the proposed LSTM estimates well the RUL of the WTs, with a Mean Absolute Error ranging between 9.40 days to 19.35 days when considering all four wind turbines. Based on these RUL prognostics, inspections are planned conservatively, well ahead of the actual day of failure. The results show that, although imperfect, prognostics are informative for maintenance and support an efficient planning of inspection tasks.

As future work we aim to improve our RUL prognostics by considering additional features such as attention mechanisms integrated into the neural networks.

Acknowledgement

The research contribution of Mihaela Mitici is partially supported as part of France 2030 program ANR-11-IDEX-0003.

References

Apunda, M. O., & Nyangoye, B. O. (2017). Challenges and opportunities of wind energy technology. *International Journal of Development Research*, *9*(06), 14174–14177.

Astolfi, D., Pandit, R., Terzi, L., & Lombardi, A. (2022). Discussion of wind turbine performance based on scada data and multiple test case analysis. *Energies*, *15*(15), 5343.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Cao, L., Qian, Z., & Pei, Y. (2018). Remaining useful life prediction of wind turbine generator bearing based on emd with an indicator. In *2018 prognostics and system health management conference (phm-chongqing)* (pp. 375–379).

Carroll, J., Koukoura, S., McDonald, A., Charalambous, A., Weiss, S., & McArthur, S. (2019). Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques. *Wind Energy*, *22*(3), 360–375.

Chen, B., Matthews, P., & Tavner, P. J. (2013). Wind turbine pitch faults prognosis using a-priori knowledge-based anfis. *Expert Systems with Applications*, *40*(17), 6863–6876.

Chen, B., Matthews, P. C., & Tavner, P. J. (2015). Automated on-line fault prognosis for wind turbine pitch systems using supervisory control and data acquisition. *IET Renewable Power Generation*, *9*(5), 503–513.

EDP. (2023). Dataset wind turbines. *https://www.edp.com/en/innovation/open-data/data*.

Elasha, F., Shanbr, S., Li, X., & Mba, D. (2019). Prognosis of a wind turbine gearbox bearing using supervised machine learning. *Sensors*, *19*(14), 3092.

Errichello, R., & Muller, J. (1994). Design requirements for wind turbine gearboxes. *NASA STI/Recon Technical Report N*, *9*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Garan, M., Tidriri, K., & Kovalenko, I. (2022). A data-centric machine learning methodology: Application on predictive maintenance of wind turbines. *Energies*, *15*(3), 826.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Graves, A., & Graves, A. (2012). Supervised sequence labelling.

Griffith, D. T., Paquette, J., Barone, M., Goupee, A. J., Fowler, M. J., Bull, D., & Owens, B. (2016). A study of rotor and platform design trade-offs for large-scale floating vertical axis wind turbines. In *Journal of physics: Conference series* (Vol. 753, p. 102003).

Herp, J., Ramezani, M. H., Bach-Andersen, M., Pedersen, N. L., & Nadimi, E. S. (2018). Bayesian state prediction of wind turbine bearing failure. *Renewable Energy*, *116*, 164–172.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Kim, K., Parthasarathy, G., Uluyol, O., Foslien, W., Sheng,

S., & Fleming, P. (2011). Use of scada data for failure detection in wind turbines. In *Energy sustainability* (Vol. 54686, pp. 2071–2079).

Kramti, S. E., Ali, J. B., Saidi, L., Sayadi, M., & Bechhoefer, E. (2018). Direct wind turbine drivetrain prognosis approach using elman neural network. In *2018 5th international conference on control, decision and information technologies (codit)* (pp. 859–864).

Kramti, S. E., Ali, J. B., Saidi, L., Sayadi, M., Bouchouicha, M., & Bechhoefer, E. (2021). A neural network approach for improved bearing prognostics of wind turbine generators. *The European Physical Journal Applied Physics*, *93*(2), 20901.

Letcher, T. (2023). *Wind energy engineering: a handbook for onshore and offshore wind turbines*. Elsevier.

Li, N., Xu, P., Lei, Y., Cai, X., & Kong, D. (2022). A self-data-driven method for remaining useful life prediction of wind turbines considering continuously varying speeds. *Mechanical Systems and Signal Processing*, *165*, 108315.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Merainani, B., Laddada, S., Bechhoefer, E., Chikh, M. A. A., & Benazzouz, D. (2022). An integrated methodology for estimating the remaining useful life of high-speed wind turbine shaft bearings with limited samples. *Renewable Energy*, *182*, 1141–1151.

Mitici, M., de Pater, I., Barros, A., & Zeng, Z. (2023). Dynamic predictive maintenance for multiple components using data-driven probabilistic rul prognostics: The case of turbofan engines. *Reliability Engineering & System Safety*, *234*, 109199.

Orozco, R., Sheng, S., & Phillips, C. (2018). Diagnostic models for wind turbine gearbox components using scada time series data. In *2018 ieee international conference on prognostics and health management (icphm)* (pp. 1–9).

Pan, Y., Hong, R., Chen, J., Singh, J., & Jia, X. (2019). Performance degradation assessment of a wind turbine gearbox based on multi-sensor data fusion. *Mechanism and machine theory*, *137*, 509–526.

Ragheb, A., & Ragheb, M. (2010). Wind turbine gearbox technologies. In *1st international nuclear & renewable energy conference (inrec)* (pp. 1–8).

Rajaoarisoa, L., Randrianandraina, R., & Sayed-Mouchaweh, M. (2024). Predictive maintenance model-based on multi-stage neural network systems for wind turbines. In *2024 international conference on artificial intelligence, computer, data sciences and applications (acdsa)* (pp. 1–7).

Rezamand, M., Kordestani, M., Carriveau, R., Ting, D. S.-K., Orchard, M. E., & Saif, M. (2020). Critical wind turbine components prognostics: A comprehensive review. *IEEE Transactions on Instrumentation and Measurement*, *69*(12), 9306–9328.

Tautz-Weinert, J., & Watson, S. J. (2017). Using scada data for wind turbine condition monitoring–a review. *IET Renewable Power Generation*, *11*(4), 382–394.

Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T. A., & Ekemb, G. (2014). Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies*, *7*(4), 2595–2630.

Teng, W., Zhang, X., Liu, Y., Kusiak, A., & Ma, Z. (2016). Prognosis of the remaining useful life of bearings in a wind turbine gearbox. *Energies*, *10*(1), 32.

Tijms, H. C. (2003). *A first course in stochastic models*. John Wiley and sons.

Tong, W. (2010). *Fundamentals of wind energy* (Vol. 44). WIT press Southampton, UK.

Wagner, H.-J. (2020). Introduction to wind energy systems. In *Epj web of conferences* (Vol. 246, p. 00004).

Yang, W., Court, R., & Jiang, J. (2013). Wind turbine condition monitoring by the approach of scada data analysis. *Renewable energy*, *53*, 365–376.

Yang, W., Tavner, P. J., Crabtree, C. J., Feng, Y., & Qiu, Y. (2014). Wind turbine condition monitoring: technical and commercial challenges. *Wind Energy*, *17*(5), 673–693.

Zaher, A., McArthur, S., Infield, D., & Patel, Y. (2009). Online wind turbine fault detection through automated scada data analysis. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, *12*(6), 574–593.

Zhao, H., Liu, H., Jin, Y., Dang, X., & Deng, W. (2021). Feature extraction for data-driven remaining useful life prediction of rolling bearings. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–10.

BIOGRAPHIES

Davide Manna holds a B.Sc. in Aerospace Engineering from University of Naples Federico II, and an M.Sc. in Aerospace Engineering from Politecnico di Torino, Italy. He has also been a visiting scholar at Utrecht University, where he conducted research on predictive maintenance of wind turbines.

Mihaela Mitici is an Assistant Professor at Faculty of Science, Utrecht University. She has a Ph.D. in Stochastic Operations Research from University of Twente, the Netherlands, and an M.Sc. in Operations Research from University of Amsterdam. During 2016-2022 she was Assistant professor at TU Delft. Mihaela specializes in Operations Research, with a focus on stochastic processes, decision-making under uncertainty, applied probability theory, machine learning. Her application domains are predictive maintenance and mobility.

Matteo D. L. Dalla Vedova is an Assistant Professor in the Department of Mechanics and Aerospace Engineering, Po-

litecnico di Torino. He holds an M.Sc. (2003) and Ph.D. (2007) in aerospace engineering from Politecnico di Torino. His research focuses on aeronautical systems engineering, particularly, the design, analysis, and numerical simulation of onboard systems, the study of secondary flight control systems, and the development of prognostics for aerospace systems. In 2017, he joined the PhotoNext laboratory of the Politecnico di Torino, working on developing and integrating optical sensors in aerospace systems.

Testing Topological Data Analysis for Condition Monitoring of Wind Turbines

Simone Casolo[1], Alexander Johannes Stasik[2], Zhenyou Zhang[3], and Signe Riemer-Sørensen[4]

[1] *Cognite AS, Oslo, Norway*
simone.casolo@cognite.com

[2, 4] *Sintef Digital, Oslo, Norway*
alexander.stasik@sintef.no
signe.riemer-sorensen@sintef.no

[3] *ANEO AS, Trondheim, Norway*
zhenyou.zhang@aneo.com

ABSTRACT

We present an investigation of how topological data analysis (TDA) can be applied to condition-based monitoring (CBM) of wind turbines for energy generation.
TDA is a branch of data analysis focusing on extracting meaningful information from complex datasets by analyzing their structure in state space and computing their underlying topological features. By representing data in a high-dimensional state space, TDA enables the identification of patterns, anomalies, and trends in the data that may not be apparent through traditional signal processing methods.
For this study, wind turbine data was acquired from a wind park in Norway via standard vibration sensors at different locations of the turbine's gearbox. Both the vibration acceleration data and its frequency spectra were recorded at infrequent intervals for a few seconds at high frequency and failure events were labelled as either gear-tooth or ball-bearing failures. The data processing and analysis are based on a pipeline where the time series data is first split into intervals and then transformed into multi-dimensional point clouds via a time-delay embedding. The shape of the point cloud is analyzed with topological methods such as persistent homology to generate topology-based key health indicators based on Betti numbers, information entropy and signal persistence. Such indicators are tested for CBM and diagnosis (fault detection) to identify faults in wind turbines and classify them accordingly. Topological indicators are shown to be an interesting alternative for failure identification and diagnosis of operational failures in wind turbines.

1. INTRODUCTION

The global demand for renewable energy sources has seen a significant rise in recent decades, with wind energy emerging as a prominent contributor to sustainable power generation (Q. Wang, Dong, Li, & Wang, 2022). Wind turbines, pivotal in harnessing wind energy, operate under diverse environmental conditions and mechanical stresses, making their maintenance and monitoring crucial for optimal performance and longevity. Condition-based monitoring (CBM) has emerged as a proactive approach to monitor the health of wind turbines, aiming to detect faults and predict potential failures before they escalate, thus minimizing downtime and maintenance costs (Stetco et al., 2019).
Traditional CBM methods often rely on spectral signal processing techniques to analyze sensor data for anomaly detection and fault diagnosis. Signal analysis techniques are commonly used for fault diagnosis and typically apply tools such as Fourier or wavelet analysis of frequency signatures from accumulated time series generated from sensors installed on wind turbines. Where possible, machine learning techniques are then used to identify early signatures of failure in the data and alert engineers as soon as the equipment's health starts deteriorating. However, frequency-based methods often require accumulating signals for a significant time before processing them successfully, making it an ideal method for analyzing failures after they occur. Online fault detection is much more challenging, and together with inherent complexity and non-linearity in wind turbine data, pose challenges for conventional analytical approaches.
To address these challenges, alternative data analysis techniques have gained attention for their ability to extract meaningful insights from complex datasets. Among those, topological data analysis (TDA) has recently risen as a possible

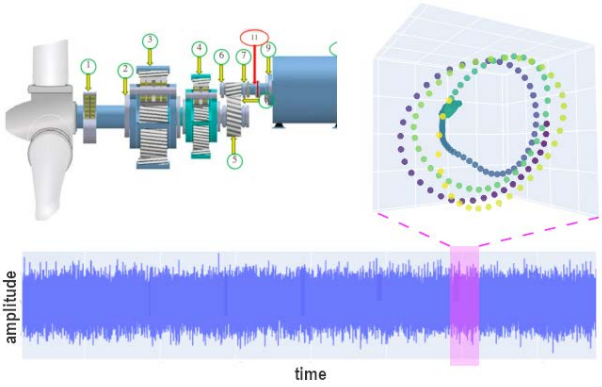


Figure 1. Overview of how the gearbox vibration data are processed by means of topological data analysis.

alternative. TDA is a branch of data analysis that focuses on revealing the the underlying structure of datasets by analyzing their shape: particularly their topology in high-dimensional state spaces. By representing data as multidimensional point clouds and leveraging mathematical tools from algebraic topology, TDA enables the identification of intricate patterns, anomalies, and trends that may not be discernible through traditional signal processing methods alone (see Fig.1).

In this study, we explore how TDA techniques can be employed to analyze vibration data collected from wind turbines at a wind park. Vibration sensors placed strategically in different locations of the turbine's gearbox provide high-frequency data capturing both vibration acceleration and frequency spectra.By employing a systematic data pipeline, including time-series segmentation and time-delay embedding, we transform the raw sensor data into a multidimensional point cloud and then, process it via topological analysis.

The primary objective of this research is to evaluate how topological indicators derived from TDA, such as Betti numbers, information entropy, and signal persistence can be used or complement more traditional spectral analysis as key health indicators for CBM and fault diagnosis in wind turbines.

2. Data description

For this analysis, we use vibration data collected from two wind turbine gearboxes from a wind park located in Norway. The data sets are proprietary, owned by the wind park operator ANEO (www.aneo.no) and this work is the first publicly available analysis of the data. The data was collected using accelerometers, located at various positions in the gearbox. For the analysis, we focused on sensors that were physically closest to the known failure positions and most correlated with the time of failure of the gearbox. The considered sensors are located at the gearbox high-speed stage front (GbxHssFr), at the gearbox intermediate stage (GbxIss), at the gearbox planetary stage (Gbx1Ps), and at the non-drive end of the generator (GnNDe). The left panel in Figure 2 show a 0.05 s example of vibrations recorded from GbxHssFr. The vibration / acceleration data were sampled at 25.6 kHz for 10 seconds at infrequent intervals. The two cases have respectively 23 and 21 samples of 10 s length with a sampling rate of 25.6 kHz. The data is collected at infrequent intervals over approximately a year until the time when failures happened, and the equipment was stopped for maintenance. In the first case, data were acquired from 2022-10-28 to 2023-10-11 and data ended with a ball bearing failure (BBF) at the non-drive end of the generator. In the second case, data was recorded from 2022-05-24 to 2023-06-21 ended with a gear tooth failure (GTF) at the planetary stage section of the gearbox.

3. Methods

In this section, we delineate the methodologies employed for analyzing complex data structures, focusing particularly on spectral analysis and topological data analysis (TDA). Spectral analysis, rooted in the principles of linear algebra and signal processing, extracts valuable insights from data by decomposing it into its constituent frequencies. Conversely, topological data analysis, drawing from the field of algebraic topology, examines the shape and connectivity of data through the lens of persistent homology, providing a holistic understanding of its underlying structure.

Both spectral analysis and TDA offer distinct yet complementary approaches to understanding complex datasets. While spectral analysis emphasizes frequency-based decomposition, TDA highlights the intrinsic topological features of the data. By comparing and contrasting these methodologies, we aim to elucidate their respective strengths, limitations, and applicability in various analytical contexts. This comparative analysis serves as a foundation for our subsequent exploration and interpretation of results, contributing to a comprehensive understanding of the dataset under investigation.

3.1. Spectral analysis

Spectral analysis, a fundamental technique in signal processing and data analysis, provides a powerful framework for decomposing complex data. Rooted in the principles of Fourier series, spectral analysis offers invaluable insights into the underlying structure and dynamics of various data types across diverse domains, including engineering, physics, biology, and finance.

At its core, spectral analysis aims to characterize the frequency content of a signal or dataset. By representing data in the frequency domain, analysts can identify dominant patterns, periodicities, and trends that may not be readily apparent in the time or spatial domain. This decomposition facilitates the extraction of meaningful information, enabling researchers to discern underlying patterns, detect anomalies, and make informed predictions.

Spectral analysis is a common tool for condition monitoring

in wind turbines (Z. Zhang, Verma, & Kusiak, 2012; Xiao et al., 2020). Vibration data are typically collected from sensors placed in correspondence to moving elements in turbine generators and gearboxes, subject to wear and mechanical failure. Data are analyzed to identify anomalies and expose drift and changes in the data that can be associated with a degradation of the system health and, inturn, lead to its mechanical failure (Tchakoua et al., 2014; Q. Wang et al., 2022; Stetco et al., 2019).

One of the key advantages of spectral analysis lies in its ability to unveil hidden relationships and structures within data. Through techniques such as Fourier transform, wavelet analysis, and singular value decomposition (SVD), analysts can disentangle complex signals into simpler components, each representing a distinct frequency or mode of variation. This spectral decomposition forms the basis for a wide range of applications, including signal filtering, noise reduction, feature extraction, and system identification.

3.2. Topological data analysis

Topological data analysis allows the interpretation of the spatial arrangement of data. This approach has been developed in the last decade and successfully applied to the analysis of data in several fields of engineering, fluid mechanics (Casolo, 2022), physics and biology (Wasserman, 2018). Here we will present a brief introduction to the topic: for a full exposition of this approach, we recommend the excellent articles from Perea and Harer (Perea & Harer, 2015), Chazal et al. (Chazal & Michel, 2021) and Smith et al. (Smith, Dłotko, & Zavala, 2021).

A common assumption in data analysis is the hypothesis that there exists a suitable space of parameters where data happen to form a manifold. In this case, it would be fair to assume that the shape of such a manifold would contain information about the data. TDA is one of the tools that can be used to interpret such information. Univariate time series of a scalar signal is not immediately suitable to be analysed with TDA. The signal is therefore embedded with a time-delay approach to form a high-dimensional space via a procedure known as Takens embedding(Takens, 1981). This method embeds a time signal into a vector without loss of information, by defining two parameters: the time-delay τ and the embedding dimension d. Then, the time series $\mathbf{x}(t)$ is sampled in d-points, each separated by a time τ. The embedded d-dimensional vector is then built as:

$$\mathbf{x}(t) = \{x(t), x(t-\tau), \ldots, x(t-d\tau)\} \tag{1}$$

As the time series evolves in time, it can be sampled repeatedly to build a series of vectors, which are accumulated to form a point cloud in d-dimensions. This cloud samples the manifold on which the data lays.

Once the data are represented in the d-dimensional space of the embedding, this can be analyzed by using algorithms developed in algebraic topology. To build the manifold, it would be required to connect each vector, *i.e.* point in the cloud within a given radius around each point, to form a network or a cell complex. This process is performed by connecting points lying within a given radius via the creation of Vietoris-Rips complexes: a simplicial (cell) complex representing the connectivity between data points in a dataset. To encode the complexity of the point cloud, we then compute a nested series of complexes that are formed at every point increasing the value of the radius in a process known as filtration. The construction of the complex involves considering all possible subsets of data points and connecting those that are within a specified distance threshold. Overall, the point cloud generated from the time series is a sampling of the shape of the data, and the filtration process generates several simplicial complexes which are the computational descriptions of the shape of the data. As the filtration parameter increases, the Vietoris-Rips complex captures increasingly complex topological features, ranging from individual points to higher–dimensional structures such as loops and voids. Typically, these features are unique to the data manifold (Attali, Lieutier, & Salinas, 2011) and are the topological structures we consider when analyzing the data.

The presence of loops, voids, etc. is encoded in the concept of homology. Persistent homology analyze the development of data sets by considering the evolution of topological features across different scales. It quantifies the persistence of these features as they emerge, merge, or disappear, providing a robust framework for capturing and characterizing the essential topological structure of complex datasets. Each structure then has a birth and a death value at a given radius of the filtration process, which can be recorded in a diagram known as a persistence diagram, unique for the analyzed shape. Each point in the diagram corresponds to a topological feature per each dimension (connected components in dimension 0, loops in dimension 1, voids in dimension 2, etc.) with its coordinates indicating the scale at which the feature is born and dies (see Figure 2 for an example of a persistence diagram). The persistence of a feature is measured as the difference between its death (d) and birth (b) scales. Naturally, persistence diagrams are non-empty only above the diagonal as the death of a feature would occur only after its birth, and the more 'persistent' a feature is, the further this would lay from the diagonal line. By analyzing persistence diagrams, it is possible to identify persistent features that are robust across multiple scales and distinguish them from transient noise or artefacts in the data. Topological indicators in each homology dimension H_k can be extracted from persistence diagrams and used to analyze data.

While TDA can be applied to uncover the shape of the data manifold for a signal of an arbitrarily long time, it can also be applied to a sequence of short time windows, sliding forward in time and partially overlapping (Perea & Harer, 2015). This sliding windows approach can be used to uncover the lo-

cal structure of data and their evolution and it has been used successfully to study the dynamics of mechanical systems.

3.3. Topology of vibration signals

Topological methods are expected to work particularly well for analyzing periodic time signals and their changes. Mathematically, it can be shown that periodic signals which can be approximated with a trigonometric function of a given frequency, can be embedded into a point cloud of elliptical shape, hence in a loop that should be detected by a high persistence signal of dimension 1 (H_1) (Perea & Harer, 2015). When the signal is instead composed of combinations of more frequencies these give rise to more complex manifolds such as tori and higher dimensional structures (Perea, 2016).
In the case of oscillating systems, according to the Arnol'd-Liouville theorem in dynamical system theory, systems of n harmonic oscillators give rise to trajectories on a n-dimensional torus. This phenomenon emerges due to the conservation of action variables, which characterize the system's motion in phase space. In a system of harmonic oscillators, each oscillator contributes a set of action-angle variables, representing the oscillation's amplitude and phase in each dimension. These variables remain constant over time, preserving the system's dynamics. As a consequence, trajectories in phase space form closed loops, tracing out toroidal surfaces (Arnol'd, 1989). This behaviour stems from the periodicity of harmonic motion, enabling the system's state to return to its initial configuration after completing a cycle. The toroidal topology of these trajectories reflects the periodicity and conservation of action variables, illustrating a fundamental principle of dynamical systems theory.
When a vibrating mechanical system such as the gearbox of a wind turbine oscillates, it is reasonable to expect, accounting for deviation and noise, a behaviour similar to that of a harmonic oscillator, hence a trajectory in phase space spanning a manifold similar to a torus. In this case, it would be reasonable to expect some homology signatures that should be visible from the persistence diagrams, making persistent homology a good candidate method for characterizing the dynamics of vibrations at the gearbox and, hopefully, spotting the appearance and evolution of abnormal behaviour from sensors' time series.

3.4. Analysis strategy

In this work, we have chunks of high-frequency data sparsely collected, each a few weeks or months apart. Every chunk of data is sampled with $25.6\,\mathrm{kHz}$ for a period of $10\,\mathrm{s}$, allowing for spectral, spectral-temporal or topological data analysis. We assume that any changes happen on time scales of days or weeks, and hence the data is stationary over each of those $10\,\mathrm{s}$ segments. Therefore, the main strategy of our analysis focuses on finding trends between time segments as we get closer to the failure time.

The key challenge in this work is the lack of ground truth, as we do not know the onset of the damage that eventually led to the failure of the gearbox. Therefore, we use the early stages of data as a baseline, assuming that the damage developed later. In other words, we are looking for systematic deviations from the early state which is assumed to be *healthy*. Topological data analysis was performed with the Giotto-TDA code suite (Pérez, Hauke, Lupo, Caorsi, & Dassatti, 2021). Time series from vibration sensors were embedded using Takens embedding with the optimal time delay and embedding dimension chosen by the built-in standard heuristics based on mutual information (Fraser & Swinney, 1986; Abarbanel, Kennel, & Brown, 1992). Persistence diagrams D were then compiled from the Vietoris-Rips complexes obtained from the filtration and used to compute the following topological indicators:
The maximum persistence, defined as the infinity norm for each homology dimension:

$$\mathcal{P}_{\infty}^{\mathrm{H}_k}(D_{\mathrm{H}_k}) = \max_{\{b,d\}\in D} |d-b| \tag{2}$$

This is a useful shape indicator as noise gives rise to points in D with a short lifetime, while relevant features of the points cloud (*e.g.* loops) are expected to have high persistence.
The normalized persistence entropy is another measure of complexity (Atienza, Gonzalez-Diaz, & Rucco, 2019; Atienza, Gonzalez-Díaz, & Soriano-Trigueros, 2020), $\overline{\mathrm{E}}_{\mathrm{H}_k}(D)$, expressed as a measure of the distribution of points along the diagram based on Shannon's entropy formula:

$$\overline{\mathrm{E}}_{\mathrm{H}_k}(D) = -\frac{1}{\log_2 \mathcal{S}(D)} \sum_{\{b,d\}\in D_{\mathrm{H}_k}} \frac{|d-b|}{\mathcal{S}(D)} \log_2\left(\frac{|d-b|}{\mathcal{S}(D)}\right)$$

where the amplitude $\mathcal{S}(D_{\mathrm{H}_k})$ for a given dimension is defined as:

$$\mathcal{S}(D_{\mathrm{H}_k}) = \sum_{\{b,d\}\in D} |d-b| \tag{3}$$

Betti curves are another informative topological indicator, which measures the amount of k-dimensional topological features *i.e.* the Betti number, β_k (Hatcher, 2002), at each value of the filtration parameter. In practice, these "count" the number of k-dimensional holes of a space: β_0 represents connected components, β_1 circles, β_2 voids, etc. As an example, for a two-dimensional circle the set of Betti numbers $\{\beta_0, \beta_1, \beta_2\}$ are $\{1, 1, 0\}$, for a filled disk $\{1, 0, 0\}$, a hollow sphere $\{1, 0, 1\}$, for a filled ball $\{1, 0, 0\}$, for a torus $\{1, 2, 1\}$, etc.
Other indicators are the **f-family of indicators** defined here, as proposed by Adcock et al. (Adcock, Carlsson, & Carlsson, 2016) and used in TDA for the anomaly detection in rotating equipment for manufacturing (Yesilli, Khasawneh, & Otto, 2022b; Khasawneh & Munch, 2016) as they combine

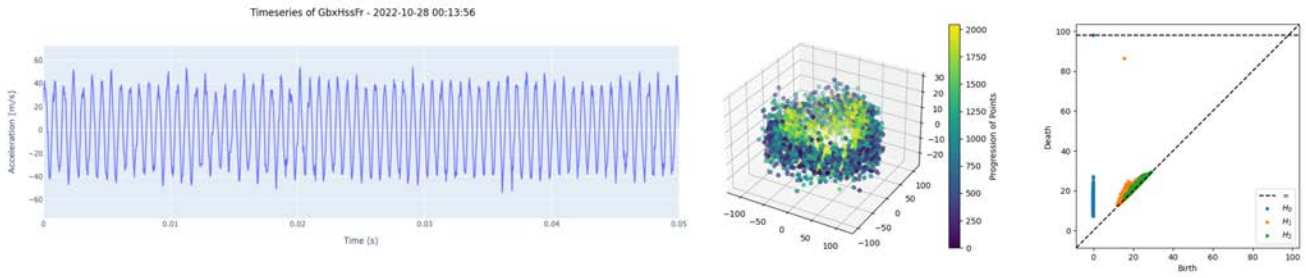


Figure 2. Left to right: Raw time-series signal, embedded point cloud and persistence diagram for GbxHssFr sensor at normal operation state. Note the toroidal point cloud, resulting from the embedding of the periodic time series. The loop structure is revealed in the persistence diagram as a point (yellow) far from the diagonal, where points created by signal noise tend to accumulate.

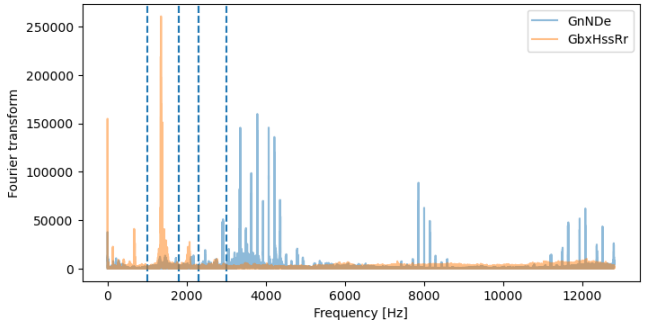


Figure 3. Fourier transform (normalised to counts) of the signal recorded on 2023-10-28 for GnNDe-BBF and GbxHssFr-BBF. The vertical lines indicate the frequency intervals for which the most dominating peaks are investigated for GbxHssFr.

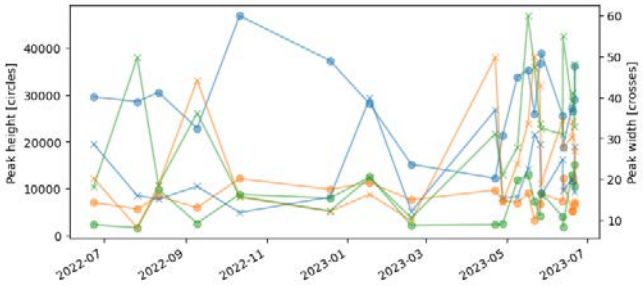


Figure 4. Peak height (left axis, circles) and width (right axis, crosses) for three frequency signatures (most dominant peak in the frequency ranges [1000, 1800], [1800, 2300], [2300, 3000] Hz) for GbxHssFr in the bearing failure case.

the highest persistence with amplitude information:

$$\begin{aligned} f_1 &= \textstyle\sum_i b_i \cdot (d_i - b_i) \\ f_2 &= \textstyle\sum_i (d_{max} - d_i) - (d_i - b_i) \\ f_3 &= \textstyle\sum_i b_i^2 \cdot (d_i - b_i)^4 \\ f_4 &= \textstyle\sum_i (d_{max} - d_i)^2 - (d_i - b_i)^4 \end{aligned} \tag{4}$$

4. Data Analysis

No data cleaning or pre-processing has been performed to the signal prior to the analysis described in Section 3, hereafter addressed as 'raw data'.

4.1. Bearing Failure

The bearing failure was reported at the non-drive end of the generator, corresponding to the location of the sensor labelled as "GnNDe" and the signal was recorded sporadically between October 2022 and the failure on November 11 2023. Each time series records acceleration data for the sensor and the corresponding frequency spectrum is computed from the raw signal through a Fast Fourier Transform (FFT) approximation. Figure 3 shows the spectrum for the signal recorded at the GnNDe (blue) and at the earliest available timestamp, 28-10-2023. We assume this to correspond to a state of "normal operations".

Topological analysis shows the point cloud corresponding with GnNDe is not describing a torus, but rather a semi-uniform ball, indicating non-periodic or very noisy behaviour. As a consequence, the H_0 persistence can only be interpreted as a measure of how much clustered or diffused the data are in the parameters space, while H_1 and higher-dimensional homology signals are expected to be low and not significant. Indeed, the only noticeable trend in the topological indicators is a decrease in H_0 persistence and an increase in entropy, typically as a consequence of a progressively less structured and more noisy signal. At a closer look, other sensor signals seem more suitable for analysis. In particular, the intermediate and high-speed stage sensors (GbxIss and GbxHss, respectively) show a more periodic and regular behaviour. Indeed the high-speed front (GbxHssFr) sensor shows a clear oscillating signal and a frequency spectrum dominated by a peak at around $1400\,\mathrm{Hz}$ and its multiples (orange spectrum in Figure 4). The embedded signal clearly shows a toroidal shape, a "filled" torus consisting of one main loop induced by the main frequency com-

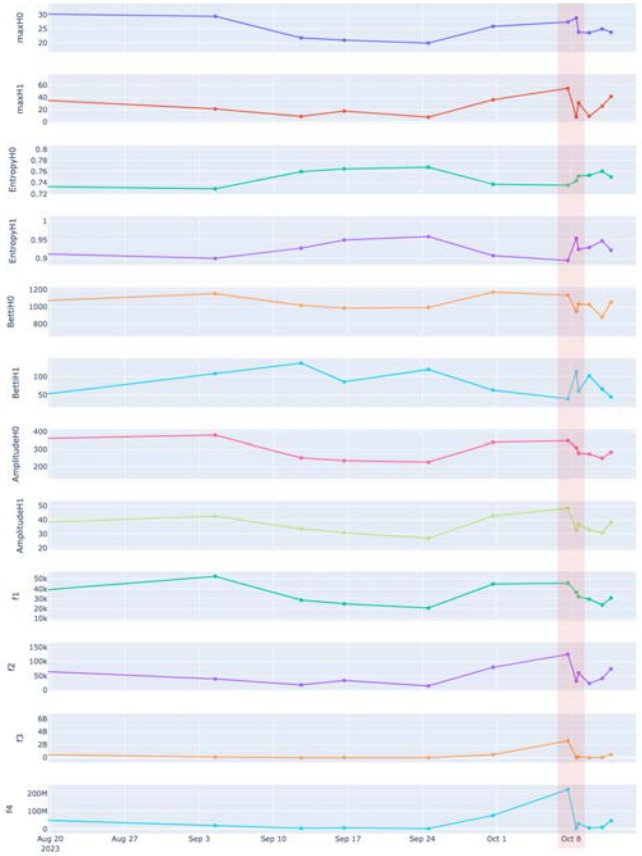


Figure 5. Topological indicators computed for the signal GbxHssFr in the bearing failure case. Highlighted the most significant anomaly, dated 2023-10-08.

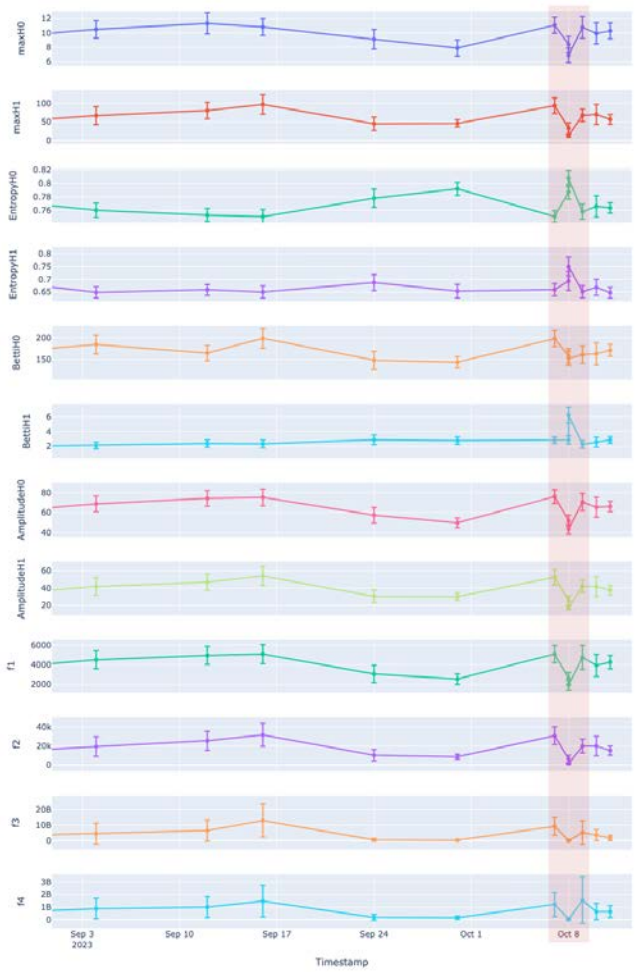


Figure 6. Topological indicators obtained by averaging the results of several sliding windows of $5\,\mathrm{ms}$, computed for each of the chunks for GbxHssFr in the bearing failure case. The most significant anomaly is dated 2023-10-08.

ponent, and the direction orthogonal to the loop blown up by the noise. The corresponding persistence diagram then shows a high persistence point for H_0 and one at H_1 corresponding to the loop and proportional to its size.

The analysis of the evolution of the GbxHssFr signal is not trivial. Figure 3 shows the time development of the most dominant peak in each of the three frequency bands shown in Figure 4. We found that the frequencies do not shift significantly until the time of the failure (not shown). The corresponding peak heights and widths show a larger spread, especially at the lowest frequency. We also measure the evolution by computing the mutual distance between the vectors containing Fourier coefficients for each time series. This distance becomes more evident between the signal in the early timestamps (i.e. normal operations) and signals in a few specific days close to the failure, in particular at 2023-10-08 and 2023-10-10, one and three days from the failure, especially for the components included from 0 to $1800\,\mathrm{Hz}$.

We observe a similar behaviour in the skewness and kurtosis of the raw signal, which show a slow decreasing trend, with a very high spike in the latter at the timestamp 08-10-2023, 3 days from the point of failure, which was not evident from the spectra alone. The monitoring of kurtosis in the early detection of bearing failures is well-known in the literature and it is likely to be a good indicator in this case as well (H. Zhang, Chen, Du, & Yan, 2016; Chauhan et al., 2024; Sawalhi & Randall, 2004).

The development of TDA indicators over time are shown in Figure 5 for GbxHssFr. Indeed most indicators show a sharp change around 08-10-2023, particularly the indicators that include the maximum persistence in dimension 1, e.g. $\mathcal{P}_{\infty}^{\mathrm{H}_1}$, $f_2(H_1)$ and $f_4(H_1)$. When applying the sliding windows approach of TDA and focusing on the short-term dynamics of the signal, the topological indicators are computed for short time windows ($5\,\mathrm{ms}$) across one signal and then averaged (Figure 6). This deep dive allows us to expose the dynamics of the signal, how the topology of the point cloud changes on short timescales and, in turn, whether the signal frequencies are finely modulated. The sliding window anal-

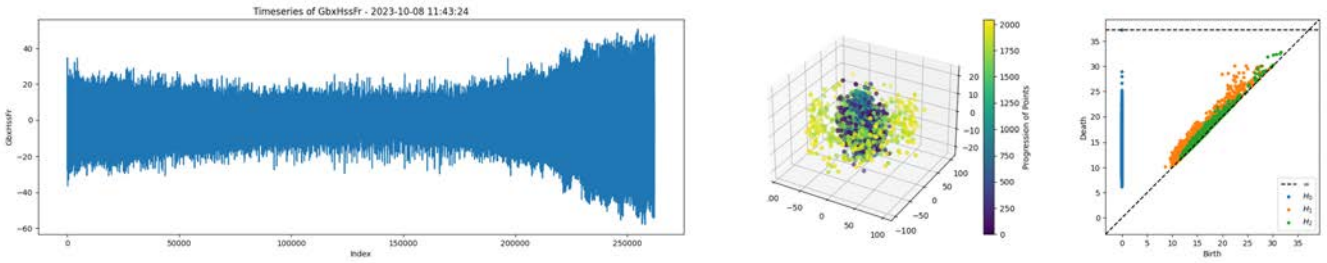


Figure 7. Left to right: Raw time-series signal, embedded point cloud and persistence diagram for GbxHssFr sensor recorded at 2023-10-08. Comparing the point and the persistence diagram with Figure 2 the loop structure of the point cloud has disappeared, together with the high persistence H_1 point in the diagram.

ysis is in perfect agreement with the Fourier analysis and the kurtosis signal, where a sharp change is visible on 2023-10-08. The change in the TDA results can be ascribed to a change in the average frequency of the signal, leading to a shrinkage of the toroidal point cloud to the point of almost closing the 'hole' of the torus (see Figure 7). This leads to a temporarily abrupt decrease in the persistence of the H_1 feature, and an increase in its entropy (entropy scales inversely with the smoothness of the manifold). There is also an apparent amplitude modulation of the raw signal which is hard to capture with TDA, but has been linked before with bearing failures in wind turbines (Jiang, Zhang, Xiang, Yu, & Xu, 2023).

4.2. Gear-tooth failure

A gear tooth damage event was reported on a different wind turbine in the same wind park in July 2023. The signal recorded for the sensor located closest to the failure, Gbx1Ps, has a frequency spectrum fairly similar to that of the high-speed sensor, GbxHssFr: dominated by few isolated frequency contributions. The only significant feature we could identify in the data is a drift in the peak width, similar to the case of the bearing fault, starting around May 2023 (see Figure 8).

Interestingly, when integrating the spectrum in the frequency range recommended by standard ISO 10816-3 (hereafter denoted Gbx1Ps.ECU2 where the signal is demodulated between 500-2kHz with the RMS broadband value between 1-150Hz). it appears more evident that a sudden jump in the signal of about 50% occurs between April and May 2023, as shown in Figure 8.

Following the same process as for the bearing fault, we focus on the high-speed gear sensor GbxHssFr, which shows a more regular oscillation pattern (see Figure 9). We apply both the Fourier and TDA analysis to uncover any possible failure signature in the data. Analogously to the bearing fault case, skewness and kurtosis show a drop, associated with an increase in the signal's median, starting from around May 2023.

The topology of the data is again that of a "filled" torus (Figure 9), which is topologically equivalent (homotopy equivalent) to a circle in 2 dimensions. This means that it should be possible to reduce the dimensionality of the Takens embedding to 2, without loss of information. We, therefore, focused on this reduced model for our analysis. The sliding windows processing for the GbxHssFr signal reveals a change in most of the topological indicators ($\mathcal{P}_{\infty}^{\mathrm{H}_{0,1}}$, $\overline{\mathrm{E}}_{\mathrm{H}_{0,1}}$, $\mathcal{S}_{H_{0,1}}$), etc.) at the same timestamp in April, and again more sharply only 2 days before the failure in July 2023, as visible from Figure 10. Close to the failure there is an increase in the persistence and a decrease in the entropy, signalling a change in the size of the loop when averaged across the 10 s of the signal at a given timestamp, but not in its shape as the Betti number indicator for dimension 1 remains stable.

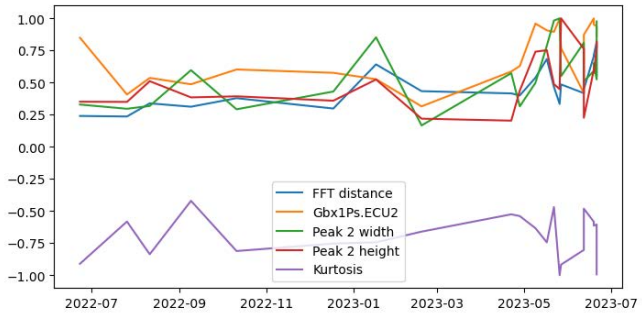


Figure 8. Selection of health indicators for the gear-tooth failure from sensor Gbx1Ps. FFT distance is the geometric distance between the average of the first three spectra in the dataset and each individual spectrum in the range [1800, 3000] Hz. Gbx1Ps.ECU2 is the indicator from standard ISO 10816-3. Peak 2 width and height are the characteristics of the dominating peak in the range [1800, 2300] Hz. Kurtosis is the kurtosis of the raw vibrations. All quantities have been normalised to their maximum value in the time interval.

In TDA, periodic functions get embedded in loops of a size proportional to the size of the sliding window (Perea & Harer, 2015), therefore, a change in the size of the torus loop should correspond to changes in the period of the gearbox vibrations or some kind of frequency modulation, close to the failure event. By looking at spectrograms for the GbxHssFr signal (Figure 11) it is possible to recover some of the dynamics of the peaks in the spectrum. On one hand, at timestamps far from the failure, the spectra shift only slightly across the 10 s of the recorded signal, and mostly the peaks tend to change

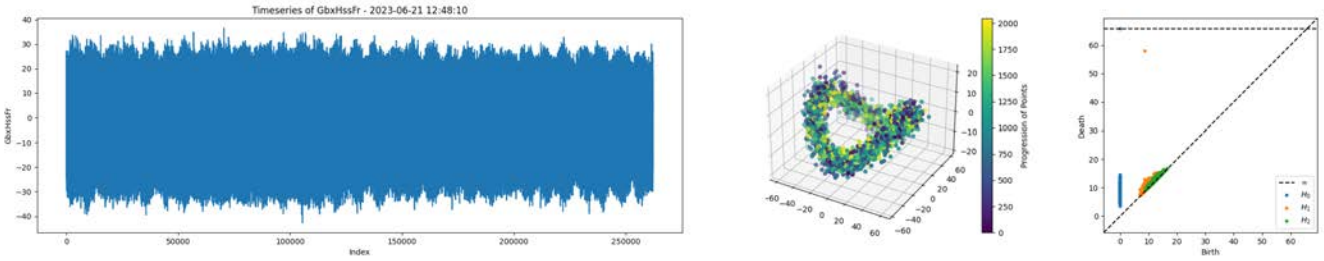


Figure 9. Left to right: Raw time-series signal, embedded point cloud and persistence diagram for GbxHssFr sensor in the gear-tooth failure case. Note the toroidal point cloud, resulting from the embedding of the periodic time series.

width with a timescale of a few seconds. On the other hand, close to the failure it appears that the two main peaks at 1350 Hz and 2700 Hz "jump" as their relative height tends to oscillate on a 3-4 Hz timescale, i.e. about 40 times across the measurement duration in a jittering fashion. This frequency modulation should also be noticeable in the TDA results, as the size of the loop in the point cloud should change as well. Indeed, this becomes evident when looking at the maximum persistence in dimension 1 ($\mathcal{P}_\infty^{\mathrm{H}_1}$) and in particular to the radius of gyration (R_{gyration}) for the point cloud, defined as:

$$R_{\text{gyration}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{\text{CM}})^2} \tag{5}$$

where N is the total number of points in the point cloud, $\mathbf{r}_i$ represents the position vector of the i-th point, $\mathbf{r}_{\text{CM}}$ denotes the position vector of the centre of mass of the point cloud.

The gyration radii for the data farther and closest in time to the gear tooth failure are shown in Figure 12 and manifests as a rapid oscillation in the gyration radius. This rapid modulation could indeed be a signature of imminent equipment failure. Interestingly, we notice this kind of modulation is common in other engineering disciplines, such as metal turning and machining, where is a signature of "chattering", a pathological resonance in the turning process (W.-K. Wang, Wan, Zhang, & Yang, 2022). Unsurprisingly, TDA has been successfully applied to chatter detection and it was shown to be useful in the early detection and the machine learning identification of such anomalies is several industrial settings (Khasawneh, Munch, & Perea, 2018; Yesilli et al., 2022b; Khasawneh & Munch, 2016; Yesilli, Khasawneh, & Otto, 2022a).

5. Conclusion

In this study, we have explored the application of topological data analysis (TDA) in conjunction with spectral analysis for condition-based monitoring (CBM) of wind turbines. Our investigation focused on analyzing vibration data aiming to detect and diagnose potential faults in gearbox components. Through TDA, we transformed raw vibration data into multi-dimensional point clouds and leveraged topological indicators such as Betti numbers, persistence diagrams, and entropy to characterize the underlying structure of the data. We compared TDA with traditional spectral analysis methods and observed that TDA offers complementary insights, particularly in identifying complex patterns and anomalies that may not be apparent through conventional signal processing techniques alone.

Our analysis revealed promising results in using TDA for fault detection and diagnosis. In the case of bearing failure, we observed significant changes in topological indicators, particularly in persistence and entropy, preceding the failure event. Similarly, for gear-tooth failure, TDA highlighted distinct changes in the structure of the point cloud, indicating the onset of damage. Furthermore, by integrating spectral analysis with TDA, we were able to uncover additional dynamics in the data, such as frequency modulation, which could serve as early indicators of equipment deterioration. These findings suggest the potential of TDA as a valuable tool for CBM in wind turbines, offering a complementary approach to monitoring and diagnosing faults and to proactive maintenance strategies in renewable energy generation. While TDA is only slightly more computationally demanding than the more traditional spectral analysis methods, it offers additional visual support by providing a manifold representing the data. Changes in the manifold of data in phase space correspond to changes in the vibration dynamics of the system, as is well known from dynamical system theory and therefore changes in the system's health may be more easily inferred by analyzing the shape of the data in addition to its spectral features.

Future research could explore the integration of TDA with machine learning techniques for more robust fault detection algorithms. Additionally, incorporating real-time monitoring capabilities could enhance the practical applicability of TDA in industrial settings.

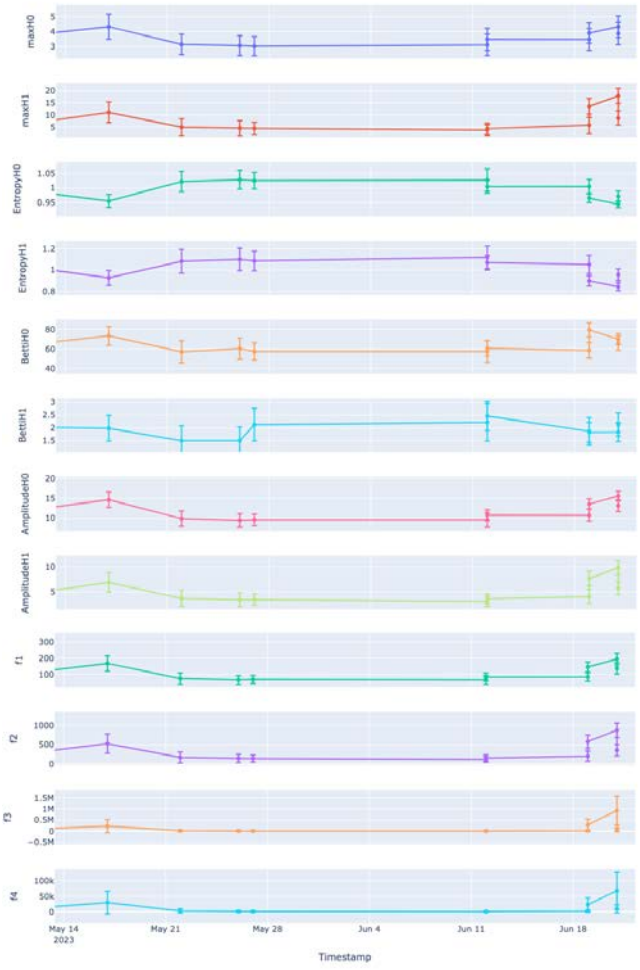


Figure 10. Topological indicators obtained by averaging the results of several sliding windows of 5 ms, computed for each of the signal GbxHssFr in the gear tooth failure case.

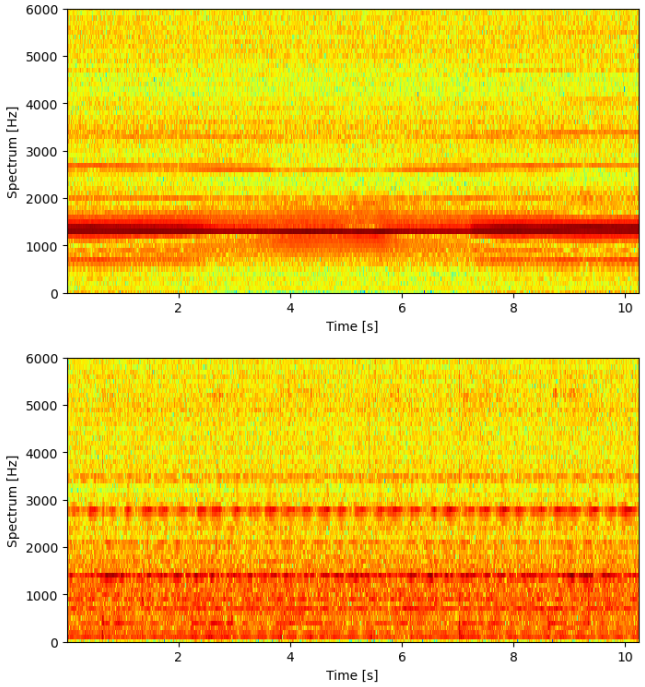


Figure 11. Spectrogram of first and second to last data point before failure.

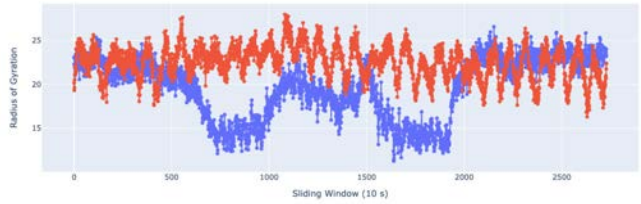


Figure 12. Radius of gyration from GbxHssFr vibration data recorded at the first data point (blue) and the last data point (red) before the failure event.

ACKNOWLEDGMENT

This publication has been funded by the SFI NorwAI, (Centre for Research-based Innovation, 309834). The authors gratefully acknowledge the financial support from the Research Council of Norway and the partners of the SFI NorwAI, in particular Aneo who shared their data.

NOMENCLATURE

Note that this section is optional.

TDA	Topological Data Analysis
CBM	Condition Based Monitoring
Gbx	Gearbox
SVD	Singular value decomposition
BBF	Ball bearing failure
GTF	Gear Tooth Failure
RMS	Root-Mean-Square

REFERENCES

Abarbanel, H., Kennel, M., & Brown, R. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, *45*, 3403–3411.

Adcock, A., Carlsson, E., & Carlsson, G. (2016). The ring of algebraic functions on persistence barcodes. In *Homology, homotopy and applications* (Vol. 18, p. 381-403).

Arnol'd, V. I. (1989). *Mathematical methods of classical mechanics*. Springer.

Atienza, N., Gonzalez-Diaz, R., & Rucco, M. (2019). Persistent entropy for separating topological features from noise in vietoris-rips complexes. *Journal of Intelligent Information Systems*, *52*, 637–655.

Atienza, N., Gonzalez-Díaz, R., & Soriano-Trigueros, M. (2020). On the stability of persistent entropy and new summary functions for topological data analysis. *Pat-*

tern Recognit., *107*, 107509.

Attali, D., Lieutier, A., & Salinas, D. (2011). Vietoris-rips complexes also provide topologically correct reconstructions of sampled shapes. In *Proceedings of the twenty-seventh annual symposium on computational geometry* (pp. 491–500).

Casolo, S. (2022). Severe slugging flow identification from topological indicators,. *Digital Chemical Engineering,*, *4*, 100045.

Chauhan, S., Vashishtha, G., Kumar, R., Zimroz, R., Gupta, M. K., & Kundu, P. (2024). An adaptive feature mode decomposition based on a novel health indicator for bearing fault diagnosis. *Measurement*, *226*, 114191.

Chazal, F., & Michel, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Front. Artif. Intell.*, *4*, 667963.

Fraser, A., & Swinney, H. (1986). Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, *33*(2), 1134–1140.

Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press.

Jiang, Z., Zhang, K., Xiang, L., Yu, G., & Xu, Y. (2023). A time-frequency spectral amplitude modulation method and its applications in rolling bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *185*, 109832.

Khasawneh, F. A., & Munch, E. (2016). Chatter detection in turning using persistent homology. *Mechanical Systems and Signal Processing*, *70-71*, 527-541.

Khasawneh, F. A., Munch, E., & Perea, J. A. (2018). Chatter classification in turning using machine learning and topological data analysis. *IFAC-PapersOnLine*, *51*(14), 195-200.

Perea, J. A. (2016). Persistent homology of toroidal sliding window embeddings. In *2016 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 6435–6439).

Perea, J. A., & Harer, J. (2015). Sliding windows and persistence: as application of topological methods to signal analysis. *Found. of Comput. Mathematics*, *15*, 799.

Pérez, J. B., Hauke, S., Lupo, U., Caorsi, M., & Dassatti, A. (2021). *giotto-ph: A Python library for high-performance computation of persistent homology of Vietoris-Rips filtrations.*

Sawalhi, N., & Randall, R. B. (2004). The application of spectral kurtosis to bearing diagnostics. *Proceedings of Acoustics*, 040115.

Smith, A. D., Dłotko, P., & Zavala, V. M. (2021). Topological data analysis: Concepts, computation, and applications in chemical engineering. *Comput. Chem. Eng.*, *146*, 107202.

Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., … Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renew. Energy*, *133*, 620-639.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand & L.-S. Young (Eds.), *Dynamical systems and turbulence, lecture notes in mathematics* (Vol. 898, p. 366-381). Springer-Verlag.

Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T. A., & Ekemb, G. (2014). Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies*, *7*, 2595-2630.

Wang, Q., Dong, Z., Li, R., & Wang, L. (2022). Renewable energy and economic growth: New insight from country risks. *Energy*, *238*, 122018.

Wang, W.-K., Wan, M., Zhang, W.-H., & Yang, Y. (2022). Chatter detection methods in the machining processes: A review. *Journal of Manufacturing Processes*, *77*, 240-259.

Wasserman, L. (2018). Fault analysis and condition monitoring of the wind turbine gearbox. *Annual Review of Statistics and its Application*, *5*(3), 501-532.

Xiao, F., Tian, C., Wait, I., Yang, Z., Still, B., & Chen, G. S. (2020). Fault analysis and condition monitoring of the wind turbine gearbox. *Advances in Mechanical Engineering*, *12*(3).

Yesilli, M. C., Khasawneh, F. A., & Otto, A. (2022a). Chatter detection in turning using machine learning and similarity measures of time series via dynamic time warping. *Journal of Manufacturing Processes*, *77*, 190-206.

Yesilli, M. C., Khasawneh, F. A., & Otto, A. (2022b). Topological feature vectors for chatter detection in turning processes. *Int J Adv Manuf Technol*, *119*, 5687–5713.

Zhang, H., Chen, X., Du, Z., & Yan, R. (2016). Kurtosis based weighted sparse model with convex optimization technique for bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *80*, 349-376.

Zhang, Z., Verma, A., & Kusiak, A. (2012). Fault analysis and condition monitoring of the wind turbine gearbox. *IEEE Transactions on Energy Conversion*, *27*(2), 526-535.

Test-Training Leakage in Evaluation of Machine Learning Algorithms for Condition-Based Maintenance

Omri Matania[1, *], Roee Cohen[1, *], Eric Bechhoefer[2], and Jacob Bortman[1]

[1]*Ben-Gurion University of the Negev, Beer Sheva, 8410501, Israel*
omrimatania@gmail.com / omrimat@post.bgu.ac.il
coroe@post.bgu.ac.il
jacbort@bgu.ac.il

[2]*GPMS International Inc., 93 Pilgrim Place, Waterbury, Vermont, 05676, USA*
eric@gpms-vt.com

ABSTRACT

Many articles have been published utilizing machine learning algorithms for condition-based maintenance through the analysis of vibration signals. One extensively researched topic is the classification of fault types in rolling bearings. There is a fairly widespread problem in the evaluation of these learning algorithms, where the separation of examples between the test and training sets is incorrect, leading to an optimistic conclusion about the algorithm's performance even when it is not the case. In this article, we will review this issue and explain how the data should be properly divided between the test and training sets to avoid this occurrence.

1. INTRODUCTION

Condition-based maintenance of rotating machinery, through the analysis of vibration signals, can significantly reduce maintenance costs and also help prevent catastrophic accidents (Matania et al., 2024; Randall, 2021). Over the years, a wide variety of machine learning algorithms have been developed to enhance traditional signal processing methods for vibration analysis (Lei, 2017).

One of the topics extensively explored in the field is the classification of fault types in bearings using machine learning algorithms (Lei et al., 2020). In this task, the algorithm is required to predict the fault type from four possibilities for a given input record: healthy condition (i.e., no fault), fault in the inner race, fault in the outer race, or fault in the rolling element. To achieve this, the algorithm is provided with examples of input records with various fault types during the training phase, and it predicts the fault type for new input records during the testing phase.

A wide variety of machine learning algorithms have been applied to this task. The first type comprises classical machine learning algorithms, where a domain expert extracts correlated features related to the fault, and the learning algorithm learns the relationship between these features and the fault type (Shalev-Shwartz & Ben-David, 2014c). The second type, developed later during the third wave of deep learning, utilizes deep neural networks to address this problem. Unlike classical algorithms, the neural network autonomously learns features that connect the vibration signals to the fault type, essentially eliminating the need for a domain expert (Goodfellow et al., 2016). In both types of learning algorithms, many studies incorrectly split the training set and the test set, leading to significant test-training leakage (Kapoor & Narayanan, 2023) that results in inaccurate, overly optimistic performance evaluations of the examined algorithms (Hendriks et al., 2022).

The first type of test-training leakage, which is also the more problematic of the two, involves splitting the same input record into different segments and randomly distributing them between the test set and the training set. Figure 1 illustrates this type of splitting. This splitting is fundamentally flawed, as many features in the same input may be unrelated to the fault type, causing the learning algorithm to inadvertently learn them. Often, when disassembling the test rig to change the tested bearing, there is a change in the vibration signature unrelated to the fault type at all. For example, researchers from the SKF group found evidence of this phenomenon in a study on fault severity assessment (Liefstingh et al., 2021). They demonstrated that the learning algorithm learned features from the vibration signature related to the transfer function of the test rig instead of information related to the fault. In such

[*]Equally contributed.

a case, when segments from the same record are divided between the training and test sets, the learning algorithm may seem to predict the fault type well, although it actually relates the segments from the same records based on the characteristics of the transfer function.

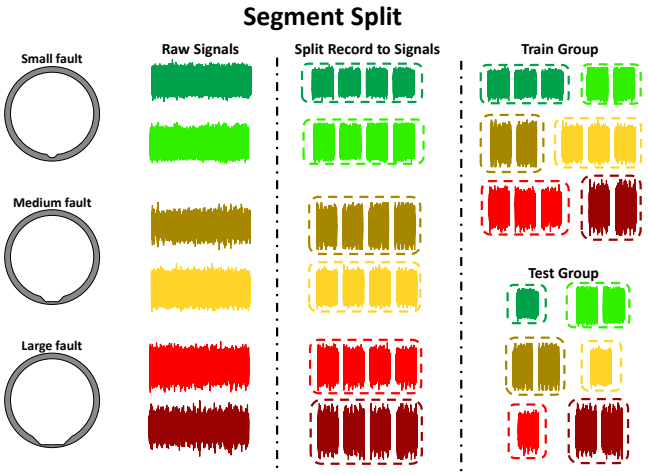


Figure 1. Illustration of random split of segments between the training and test sets.

The second type of test-training leakage involves the random separation of different records of the same fault type with the same fault shape precisely between the test and training sets. Figure 2 illustrates this type of improper separation. Each fault type can exhibit a wide variety of shapes. For example, a fault in the outer race can manifest in numerous different shapes and sizes, potentially even an infinite number. In practice, the likelihood that the exact shape of the fault in a real-world scenario matches one of the faults the algorithm learned from in the training set is very low. Many datasets record each fault multiple times. Randomly distributing these records between the test and training sets is incorrect and does not represent reality. In such a scenario, the algorithm may learn features related to the shape of the fault rather than its type, leading to overly optimistic evaluated performances. Furthermore, in some cases, the records of the same fault shape do not include the assembly of the test rig. Consequently, the algorithm may learn features from the vibration signature related to the transfer function of the test rig rather than information related to the fault, similar to the previous case of random segment split.

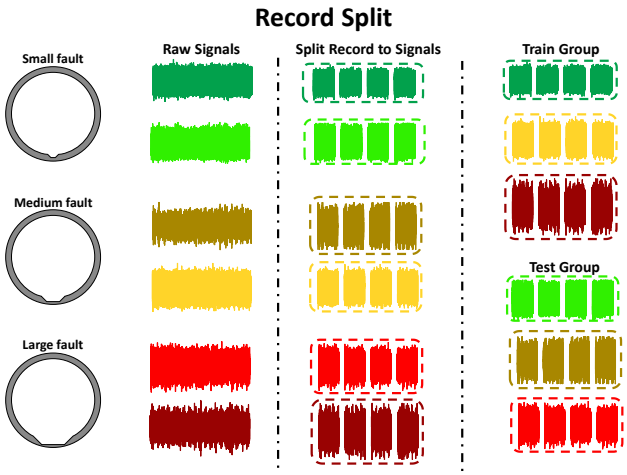


Figure 2. Illustration of random split of records between the training and test sets.

Figure 3 illustrates the correct splitting for evaluation learning algorithms: all records of each fault shape are either sent to the test set or to the training set. Following this separation, each record can be further divided into smaller segments if necessary. In this approach, to achieve an accurate estimation of performance, it is recommended to use K-fold testing. For example, in this study, performance evaluation in the test is implemented using the leave-one-out procedure, which is an extreme form of K-fold testing.

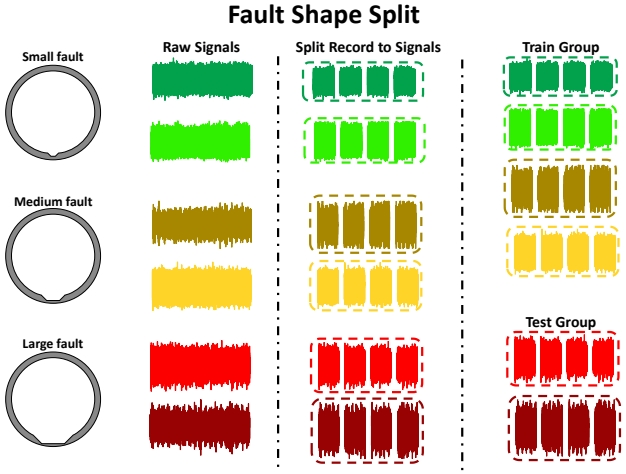


Figure 3. Illustration of split by fault shape between the training and test sets.

Section 2 will discuss the datasets analyzed in the article, and Section 3 will cover the learning algorithms. Section 4 will demonstrate that indeed, both segment split and record split lead to optimistic results compared to the correct way of fault shape split. Section 5 will summarize the article and present the conclusions.

2. TESTED DATASETS

Two datasets that are frequently used for evaluating machine learning algorithms for fault classification in rotating

machinery are discussed in the article. The first dataset, Case Western Reserve University dataset (CWRU), is accessible via the link (*Case Western Reserve University Bearing Data Center Website*, n.d.) and is extensively described in the work by Smith and Randall (Smith & Randall, 2015). It is important to note that this dataset has several issues, as explained by Smith and Randall, yet for unknown reasons it is still widely utilized. The CWRU test rig is illustrated in Figure 4. The CWRU dataset comprises a total of 416 distinct records, but in practice, only 12 truly different faults exist.

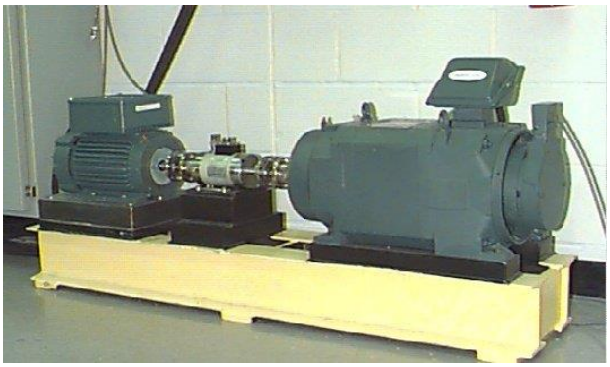


Figure 4. CWRU dataset test rig. Reproduced from (*Case Western Reserve University Bearing Data Center Website*, n.d.).

The Paderborn University (PU) dataset also serves for the evaluation of various learning algorithms and is extensively described in the study of Lessmeier et al. (Lessmeier et al., 2016). Our review of this dataset led to the conclusion that it also has several issues, such as unclear sources of interferences in the spectrum. In total, the PU dataset contains 2493 recordings, with 26 truly distinct faults in practice. Figure 5 depicts the experimental setup.

Figure 5. PU dataset test rig. Reproduced from (Lessmeier et al., 2016).

3. TESTED ALGORITHMS

In the current section, two learning algorithms will be described, which are used to demonstrate the effect of test-training leakage. The first is K-nearest neighbors (KNN) (Shalev-Shwartz & Ben-David, 2014a) and the second is Random Forest (Shalev-Shwartz & Ben-David, 2014b). All tested algorithms used the following features: mean, variance, kurtosis and absolute mean.

KNN operates by determining the class of a data point based on the majority class among its k-nearest neighbors within the feature space. The algorithm computes the distance between the given data point and its neighbors. The parameter K, denoting the number of neighbors taken into account, is a crucial factor that can significantly influence the model's performance. Small K values may result in overfitting, while large K values may lead to inadequate fitting of the training data. In the current study, K was set to 1 to prevent additional issues with training-validation splitting. Figure 6 provides a visualization of the KNN process for classification.

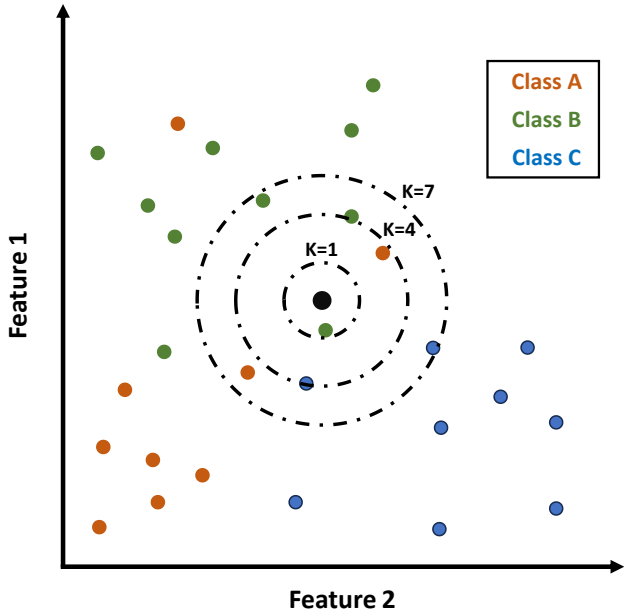


Figure 6. Illustration of KNN.

Random Forest stands out as a robust ensemble learning algorithm widely applied in machine learning for classification tasks. It generates numerous decision trees during training and outputs the mode of the classes. The key innovation of Random Forest lies in its incorporation of randomness—each tree is trained on a random subset of the data, and during each split, a random subset of features is taken into consideration. This randomness aids in mitigating overfitting and enhancing the model's generalization performance. Furthermore, for classification, the predictions from multiple trees are consolidated through majority voting, resulting in a resilient and accurate final prediction. In the current case, the number of trees was set to 300. This is a standard number of trees intended to prevent overfitting. Once again, this parameter was not set based on the validation set to avoid additional issues with training-validation splitting. Figure 7 provides a visualization of the random forest process for classification.

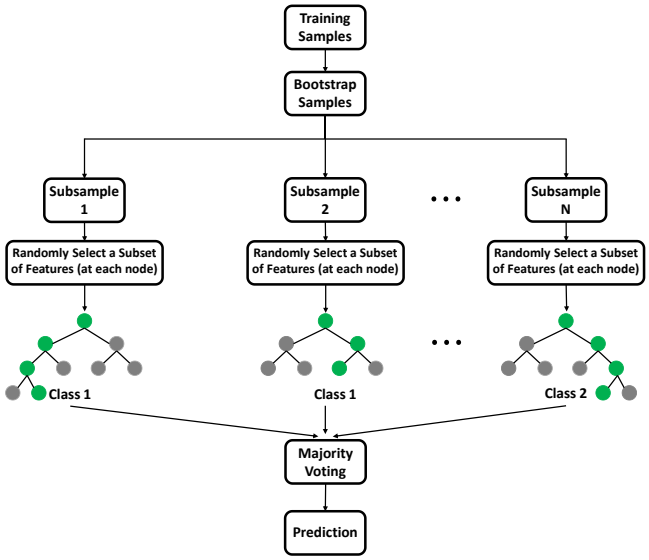


Figure 7. Illustration of random forest.

4. RESULTS

The results of KNN and random forest on the two tested datasets, CWRU and PU, are depicted in Figure 8 for the three types of splitting: segment split, record split, and fault shape split. The results of the segment and record splits were determined using a 10-fold cross-validation technique to calculate the average accuracy. The fault shape split results were obtained through a leave-one-out procedure. Furthermore, to compare the performance with a degenerated algorithm, two lines were added to the figure representing predictions of the test examples in the fault shape splitting for CWRU and PU, based on the most prevalent label in the training set. This degenerated algorithm disregards the features and, for any new unseen examples, returns the mode of the classes from the training set.

As can be seen from the figure, for the CWRU dataset, when changing from segment split to record split, the accuracy significantly decreases. For both datasets, when the correct splitting method is utilized, namely the fault shape split, the results are significantly worse. In the case of CWRU, they are even lower than the accuracy of the degenerated algorithm, which predicts the training mode constantly.

These results demonstrate that incorrect random splitting leads to overly optimistic conclusions. For the CWRU dataset, based on segment split, it seems that the very straightforward approach of using simple signal features and classic machine learning algorithms like KNN and random forest enable achieving good accuracy, close to 90%. However, when the record split is applied, the results are much less optimistic, and when the correct method is applied, the results are worser than constantly predicting the training mode, indicating that both algorithms probably learn nothing related to the fault type. For PU datasets, even when record split is utilized, the results are still optimistic, and only when the correct splitting of fault shape is utilized can we again conclude that the algorithm did not learn too much information related to the fault type.

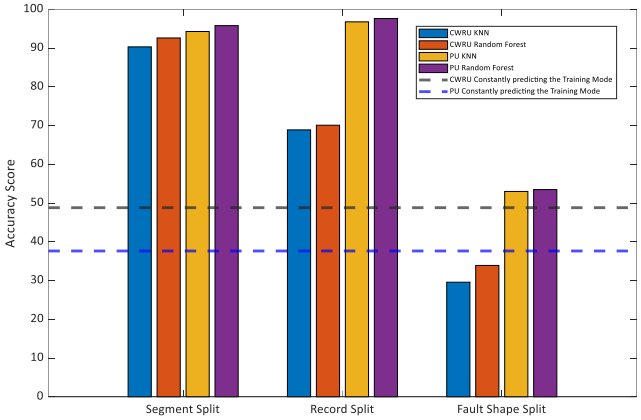


Figure 8. Accuracy score for bearing fault type classification on CWRU and PU datasets by KNN and random forest for different split approaches.

5. CONCLUSION

Many machine learning algorithms have been suggested for vibration analysis of rotating machinery for condition-based maintenance. As demonstrated in this paper, improper splitting of data between the training and test sets may lead to test-training leakage and, consequently, to an overly optimistic evaluation of the machine learning algorithm performances.

In the current study, this problem was tested on the prevalent task of fault type classification in rolling bearings. It was shown that when improper segment splitting is utilized, overly optimistic conclusions can be drawn regarding a simple approach that combines straightforward signal features with basic machine learning algorithms, as they achieve accuracy close to 90%. However, when the right splitting is utilized, reflecting the real scenario in which records of the exact same fault shape should not be present in both the training and test sets, the results are very poor and, in some cases, worser than constantly predicting the training mode, indicating that the algorithms have not learned anything.

Three further comments regarding machine learning studies in the vibration analysis field are worth discussing. First, most of the currently available datasets, such as CWRU and PU, contain many contaminated records. The research community would benefit greatly from newer datasets without contaminated records, which would also encompass a broader range of fault shapes. Second, it is not clear why so many papers attempt to solve the problem of fault type classification in bearings, as classic approaches in signal processing are adept at solving it (Randall & Antoni, 2011). We recommend that future papers focus on addressing fault

severity and estimating remaining useful life tasks (Matania et al., 2023), or alternatively, focus on fault classification of components that currently lack well-established classic approaches. Another option is to examine cases where the signal-to-noise ratio is so low that signal processing algorithms are unable to classify the fault type. The last comment worth noting is that a maintainer or operations manager doesn't really care if a bearing has a ball, inner, or outer race fault – as they will probably replace the entire bearing regardless. The more important issue is fault detection, determining whether the bearing is healthy or not. Fault classification is more interesting if it helps to better estimate severity or remaining useful life.

ACKNOWLEDGEMENT

Omri Matania is supported by the Adams Fellowships Program of the Israel Academy of Sciences and Humanities.

REFERENCES

Case Western Reserve University Bearing Data Center Website. (n.d.). Retrieved November 23, 2022, from https://engineering.case.edu/bearingdatacenter/welcome

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, *169*, 108732. https://doi.org/10.1016/J.YMSSP.2021.108732

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/J.PATTER.2023.100804

Lei, Y. (2017). Intelligent fault diagnosis and remaining useful life prediction of rotating machinery. In *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery* (1st ed.). Butterworth-Heinemann. https://doi.org/10.1016/C2016-0-00367-4

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, *138*. https://doi.org/10.1016/j.ymssp.2019.106587

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Society European Conference*, *3*(1). https://doi.org/10.36001/PHME.2016.V3I1.1577

Liefstingh, M., Taal, C., Restrepo, S. E., & Azarfar, A. (2021). Interpretation of Deep Learning Models in Bearing Fault Diagnosis. *Annual Conference of the PHM Society*, *13*(1). https://doi.org/10.36001/PHMCONF.2021.V13I1.3047

Matania, O., Bachar, L., Bechhoefer, E., & Bortman, J. (2024). Signal Processing for the Condition-Based Maintenance of Rotating Machines via Vibration Analysis: A Tutorial. *Sensors 2024, Vol. 24, Page 454*, *24*(2), 454. https://doi.org/10.3390/S24020454

Matania, O., Bachar, L., Khemani, V., Das, D., Azarian, M. H., & Bortman, J. (2023). One-fault-shot learning for fault severity estimation of gears that addresses differences between simulation and experimental signals and transfer function effects. *Advanced Engineering Informatics*, *56*, 101945. https://doi.org/10.1016/J.AEI.2023.101945

Randall, R. B. (2021). *Vibration-based condition monitoring : industrial, automotive and aerospace applications* (2nd ed.). WILEY. https://www.wiley.com/en-sg/Vibration+based+Condition+Monitoring%3A+Industrial%2C+Automotive+and+Aerospace+Applications%2C+2nd+Edition-p-9781119477556

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, *25*(2), 485–520. https://doi.org/10.1016/J.YMSSP.2010.07.017

Shalev-Shwartz, S., & Ben-David, S. (2014a). Chapter 19 - Nearest Neighbor. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057135, pp. 258–267). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shalev-Shwartz, S., & Ben-David, S. (2014b). Section 18.3 - Random Forests. *Understanding Machine Learning: From Theory to Algorithms*, *9781107057135*, 255–256. https://doi.org/10.1017/CBO9781107298019

Shalev-Shwartz, S., & Ben-David, S. (2014c). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, *64–65*, 100–131. https://doi.org/10.1016/J.YMSSP.2015.04.021

BIOGRAPHIES

Omri Matania is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Omri is a Talpiot graduate and served nine years in IDF in several roles including algorithm section

leader. He completed with honors his bachelor's degree in mathematics and physics in the Hebrew University of Jerusalem and completed his master's degree with honors in mechanical engineering in Ben-Gurion University of the Negev.

Roee Cohen is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Roee completed with honors his bachelor's degree and master's degree in mechanical engineering in Ben-Gurion University of the Negev.

Eric Bechhoefer received his bachelor's degree in biology from the University of Michigan, his master's degree in operations research from the naval postgraduate school, and a Ph.D. in general engineering from Kennedy Western University. He is a former naval aviator who has worked extensively on condition-based maintenance, rotor track and balance, vibration analysis of rotating machinery, and fault detection in electronic systems. Dr. Bechhoefer is a fellow of the prognostics health management society, a fellow of the society for machinery fault prevention technology, and a senior member of the IEEE reliability society. Additionally, Dr. Bechhoefer is also a member of the SAE committee covering integrated vehicle health management, and a member of the MSG-3, rotorcraft maintenance programs industry group.

Jacob Bortman is currently a full Professor in the department of mechanical engineering and the head of the PHM Lab in Ben-Gurion University of the Negev. Retired from the Israeli air force as brigadier general after 30 years of service with the last position of the head of material directorate. Chairman and member of several boards: director of business development of Odysight Ltd, Chairman of the board of directors, Selfly Ltd., board member of Augmentum Ltd., board member of Harel finance holdings Ltd., Chairman of the board of directors, Ilumigyn Ltd. Editorial board member of: "Journal of Mechanical Science and Technology Advances (Springer, Quarterly issue)". Head of the Israeli organization for PHM, IACMM - Israel Association for Comp. Methods in Mechanics, ISIG - Israel Structural Integrity Group, ESIS - European Structural Integrity Society. Received the Israel National Defense prize for leading with IAI strategic development program, Outstanding lecturer in BGU, The Israeli prime minister national prize for excellency and quality in the public service - First place in Israel. Over 80 refereed articles in scientific journals and in international conference.

Timeseries Feature Extraction for Dataset Creation in Prognostic Health Management: A Case Study in Steel Manufacturing

Thanos Kontogiannis[1], Wanda Melfo[2], Nick Eleftheroglou[3], and Dimitrios Zarouchas[4]

[1,3] *Intelligent and Sustainable Prognostics Group, Aerospace Structures and Materials Department, Faculty of Aerospace Engineering, TU Delft, Delft, 2629 HS, Netherlands*
a.kontogiannis@tudelft.nl
n.eleftheroglou@tudelft.nl

[1,3,4] *Center of Excellence in AI for Structures, Aerospace Structures and Materials Department, Faculty of Aerospace Engineering, TU Delft, Delft, 2629 HS, Netherlands*
d.zarouchas@tudelft.nl

[2] *Research and Development, Tata Steel Europe, IJmuiden, 1970 CA, Netherlands*
wanda.melfo@tatasteeleurope.com

ABSTRACT

This study focuses on a critical aspect of implementing prognostics and health management (PHM) for assets: the creation of a descriptive dataset. In real-world applications, dealing with sparse and unlabelled big data is common, particularly in industries like production lines where complex subprocesses are monitored by multiple sensors. Moreover, selective application of quality control means that much of the data lacks information about end properties, making datasets provided by manufacturers unsuitable for PHM frameworks. This work aims to bridge the gap between raw production data and PHM frameworks, focusing on steel manufacturing management. In the context of steel manufacturing, compromised surface quality, characterized by thicker oxide layers chipping during milling, has been observed. We propose inferring compromised coils by analyzing temperature profiles directly before the coiling station to address this. Deviations from the goal temperature profile can indicate compromised surface quality, eliminating the need for tedious oxide layer thickness measurements, which are not feasible for continuous hot strip milling processes. The available dataset comprised multiple years of production, with no direct indication of the surface quality. Exploratory clustering analysis was the first step in the lack of labels. Even though indicative of the underlying pattern of the healthy/damaged coils distinction, three shortcomings were identified. Clustering was solely based on the similarity between the temperature profiles of the coils, so no domain knowledge was included regarding the goal temperature profile. Additionally, since different steel grades have different goal profiles, the model needs to be specifically trained for each grade. Also, a soft classification between healthy and damaged can provide more detailed information about the surface quality. Coils with low-confidence classifications can be identified and treated accordingly, thereby improving PHM framework performance by providing a dataset with only high-confidence samples. To tackle these issues, an expert-knowledge-based normalization technique and feature engineering, paired with synthetic labelling, contributed to the creation of a soft neural network classifier. This study presents the reality of handling real-world data for PHM applications and highlights the need for careful and informed feature extraction. This ensures the seamless integration of PHM frameworks into real-world systems, ultimately enhancing production yield by improving end-product quality.

1. INTRODUCTION

The 4^{th} industrial revolution led to a skyrocketing increase in the available data in production and manufacturing lines. Sensors were developed and installed throughout the processes, and computer-operated regulating devices were retrofitted to production equipment. This not only meant that the manufacturing process could be guided by preset rules that were constantly tailored to the real measurements of the system but also that an enormous amount of data became available. Manufacturers, suspecting these data's value, made sure to gather and store them in databases. However, the vast majority of the available data are unstructured and unlabelled, leading to their under-utilization.

The unstructured nature of big data gathered from multiple sources during production is one of the main challenges of applying prognostic health management (PHM) frameworks (Zio, 2022). Data quality greatly affects performance, especially for fault detection (FD), which is usually the first task during a PHM framework. The need to identify anomalies and deviations from the normal operating condition of the asset arises, but data are usually high-dimensional with non-smooth distribution densities. This makes their reconstruction and, in turn, the distinction between healthy and abnormal, challenging.

In order to tackle the high-dimensionality of the data, the challenge of informative feature extraction (FE) arises (Jardine, Lin, & Banjevic, 2006), in the hope of projecting the data into manifolds of lower dimensionalities, where the underlying classes become distinguishable. Traditional pre-processing techniques include statistical feature extraction in the time domain (Caesarendra & Tjahjowidodo, 2017), fast fourier transform (Z. Wang, McConnell, Balog, & Johnson, 2014), discrete wavelet transform (Z. Wang et al., 2014), continuous wavelet transform (Kankar, Sharma, & Harsha, 2011), morphology operators (Gush et al., 2018) and principal component analysis (Choi, Lee, & Lee, 2005). Adding to that, in the latest years, due to the ever-increasing applications of machine learning (ML) and deep learning (DL) in computer science, numerous successful applications for FE for PHM frameworks have been demonstrated. Categorical adversarial autoencoders (Liu et al., 2018), stacked autoencoders (Y. Wang, Yang, et al., 2020), generative adversarial networks (Jiang, Hong, Zhou, He, & Cheng, 2019; Xia et al., 2022), deep convolutional networks (Wu & Zhao, 2018) and deep belief networks (Y. Wang, Pan, Yuan, Yang, & Gui, 2020), are examples of ML and DL frameworks to extract lower dimensionality representations of big data.

However, it becomes apparent that DL does not provide a universal solution to FE (Zhao et al., 2019). Great effort and resources are associated with designing and training a successful DL model, and usually, the impressively performing but complex architectures make the DL networks task- and domain-specific. In this work, presented with a big real-world dataset from steel manufacturing, a data science and fundamental approach for feature extraction is followed. The aim is to showcase that even highly complex datasets with high variability, can be handled with expert-driven analysis, proving the discriminating power of informative features. The following sections will describe the issue under consideration and the available dataset (Section 2), followed by an overview of the applied methods (Section 3), the findings (Section 4), and a concluding discussion (Section 5).

2. Problem statement and dataset description

Steel strips are being widely used for numerous applications across multiple domains, such as the automotive industry, the aerospace industry, chemical equipment and light manufacturing, all of which, among others, have increasing demands considering surface quality. However, surface defects can appear during manufacturing, which significantly diminishes the end surface quality of the manufactured steel strips. Known root causes of surface quality defects are material defects, process defects and corrosion defects (Z. Wang, Wang, & Chen, 2020). Material and process defects can be more easily mitigated by tailoring the material's composition and manufacturing process (i.e. rolling forces, timely inspection and replacement of rollers). Unfortunately, corrosion defects are, by nature, more challenging. The low stability of the typical three-layer oxide composition of steel (hematite Fe_2O_3, magnetite Fe_3O_4 and wustite $Fe_{1-y}O$) at the low coiling temperatures, the presence of other elements in low-carbon steel, the presence of inclusions, the continuous cooling conditions, the temperature gradient across the width of the strip, the absence or lack of oxygen in the centre regions, all affect the oxide evolution (Chen & Yuen, 2001; Deng et al., 2017). The extensive study of Min K. et. al. (Min, Kim, Kim, & Lee, 2012) revealed a correlation between the thickness of the oxide layer and the surface quality. This is attributed to the fact that a thicker oxide scale is more brittle and, thus, more prone to chip off.

As demonstrated by Min K. et al. (Min et al., 2012), measuring the oxide layer thickness during production is not feasible. Production must be halted, and the oxide layer formation must be frozen (i.e., by spraying molten glass on the surface). This process can quickly become costly and counterproductive for a real-world application. This fact, combined with the fact that practical scale differs from lab-grown (Deng et al., 2017), led our team to try to develop a way to infer it indirectly from production measurements.

The first step towards achieving this goal is creating a labelled dataset from historical data containing coils with deteriorated and pristine surface quality. We theorize that, by observing the steel strip's coiling temperature (CT) profile, major deviations from the goal temperature and, more importantly, rapid fluctuations, can indicate a chipped-off oxide layer. The reasoning behind this is that when the oxide layer chips off, some parts of exposed steel appear on the surface that have drastically different emissivity than the oxides, throwing off the pyrometer temperature measurements. Thus, the need to distinguish faulty cases from normal ones from sequential data arises.

The dataset in hand consists of the process parameters, the CT profiles and the material properties of the manufactured steel strips from the hot strip milling (HSM) process of Tata Steel Europe ©. Due to the great variability in the CT profiles as well as the goal temperatures, a single steel grade was chosen, considering its observed troublesome behaviour during milling (the details of which will not be disclosed due to confidentiality). After data cleaning, the remaining dataset consists of 3768 CT profiles, that will in turn, be used for the development of the classification algorithm.

3. METHODS

Given the dataset's unlabelled nature, an exploratory clustering analysis was the first step towards processing the dataset to discover the expected underlying pattern of healthy and damaged coils. Afterwards, a domain-specific normalization was introduced to the sequential data to assist towards creating a universal framework independent of the steel grade. This is of high importance since a great number of different steel grades are produced. Therefore, if the developed framework is grade-specific, it will need to be trained for each grade specifically, making it counter-productive. Two different FE techniques were realized and contrasted: a domain-agnostic one and an expert-knowledge-based one. Finally, synthetic labels were created to facilitate the training of a neural network (NN) soft classifier to discriminate the produced coils into healthy and damaged ones (meaning with chipped-off oxides and pristine surface quality, respectively).

3.1. K-means with Dynamic Time Wrapping

Clustering analysis is one of the first steps in processing unlabelled data, due to its ability to uncover underlying patterns and connections in the dataset without requiring any prior knowledge. A well-known and established clustering algorithm is the k-means algorithm (Lloyd, 1982). The k-means algorithm strives to partition the n available observations into k clusters, where each observation belongs to the cluster with the nearest mean, referred to as the centroids. The original algorithm works by minimizing the squared Euclidean distances between the centroids and the observations. An immediate issue can be observed when the algorithm is tasked with clustering sequential data. The Euclidean distance between two points A and B can be calculated by Eq. (1), with δ being the distance between the elements.

$$D(A,B) = \sqrt{\delta(a_1,b_1)^2 + \cdots + \delta(a_T,b_T)^2} \qquad (1)$$

If A and B are sequences with $A = \langle x,y,x,x\rangle$ and $B = \langle x,x,y,x\rangle$, their Euclidean distance according to Eq. (1), will be great, even though intuitively, they sequences are similar. This is attributed to the inability of the Euclidean distance to capture similarities that are shifted in time. For that reason, the dynamic time wrapping (DTW) metric is introduced. Its main attribute is that it can capture similarities between sequences independently of the velocity (Sakoe, 1971). The way to achieve this is by aligning the coordinates inside the sequences by minimizing Eq. (2), where A_i is the subsequence $\langle a_1,\ldots,a_i\rangle$.

$$D(A,B) = \delta(a_i,b_i) + \min\left\{\begin{matrix} D(A_{i-1},B_{j-1}) \\ D(A_i,B_{j-1}) \\ D(A_{i-1},B_j)\end{matrix}\right\} \qquad (2)$$

Even though DTW can effectively find the optimal alignment between sequences and provide a single score for similarity, k-means requires the calculation of a cluster prototype (the centroid), which is the average of the assigned observations. Petitjean et al. (Petitjean, Ketterlin, & Gançarski, 2011), proposed the DTW barycenter averaging (DBA) algorithm, which iteratively calculates the barycenter of a set of sequences for the k-means algorithm. Later on, a differentiable function for computing the soft minimum of all of the alignment costs increases performance and reduces arithmetic complexity, referred to as soft-dtw algorithm (Cuturi & Blondel, 2018). For the aforementioned reasons and considering the large size of the dataset, the soft-dtw algorithm is chosen.

3.2. Domain-specific Normalization

One of the main issues with the given application for the distinction between good and bad coils is the great difference between the goal CT profiles for different steel grades. The difference lies not only in the temperature but also in the shape of the wanted goal CT profile. Some steel grades require a coffin-shaped CT profile where the head and the tail of the coil are hotter than the middle section. Adding to that, steel strips that belong to each coil are not manufactured equally. The manufacturer provides a range of properties for each grade, and thus, the final goal CT profile depends on the exact needs of the specific order. This inevitably leads to the inability to generalize any realized framework since it would need to be redesigned and explicitly trained for each available steel grade. The authors try to alleviate this dependency by normalizing the CT profiles with the goal CT. Let $traj_1 = \langle t_1,t_2,\ldots,t_F\rangle$ and its respective goal CT $goal_{ct} = \langle g_1,g_2,\ldots,g_F\rangle$. The normalized CT profile is calculated with the following:

$$traj_1{}^j_{norm} = \frac{t_j - g_j}{g_j}, j = [1,F] \qquad (3)$$

For the remaining of the analysis, the normalized trajectories $traj_{norm}$ will be used.

3.3. Feature Extraction

For the good/bad coil distinction, a NN classifier will be utilized (as explained in Sec. 3.5). NNs are generally unable to handle and interpret sequential data as inputs, excluding recurrent NNs (RNN) (Amari, 1972). RNNs come with their own set of limitations with lengthy sequential data, namely the high computational complexity, the vanishing gradient problem and the often-required tedious hyper-parameter tuning. To partially tackle said limitations, one can choose to split the sequential data into overlapping windows, but the choice of the length and overlap of the windows adds to the complexity of choosing optimal hyperparameters. For the aforementioned reasons, traditional fully connected layers (FC) will be used, and thus, the CT profiles need to be represented with fea-

tures. Two different techniques for FE are being contrasted: A domain-agnostic approach where a plethora of features is being extracted (statistical, temporal and spectral) and filtered automatically, and an expert-knowledge-based one.

3.3.1. Domain-Agnostic FE

Given sequential data, a domain-agnostic FE refers to the process where a variety of features are extracted without considering the nature of the data in the hope of capturing as many characteristics as possible. In this study, the features extracted were:

- Statistical: max, absolute max, min, kurtosis, standard deviation, variation, mean, median, min, quantile, sum of values, length, variance, variation coefficient, count of values above/below mean value, first location of min and max, length of longest strike above/below mean, root mean square, sum of reoccurring values,
- Autocorrelation values (Yentes et al., 2013) for $lag = (1, 2, \ldots, 10)$ and descriptive statistics on the aggregation function (mean, variance, median, standard deviation) over the autocorrelation,
- Approximate entropy (Yentes et al., 2013) with $(m = 2, r = 0.1), (m = 2, r = 0.3), (m = 2, r = 0.5), (m = 2, r = 0.7), (m = 2, r = 0.9)$ with m the length of the compared run of data and r the filtering level,
- Non-linearity measure with c3 statistics (Schreiber & Schmitz, 1997) with $lag = (1, 2, 3)$,
- Complexity-invariant distance (CID) with and without normalization (Batista, Keogh, Tataw, & De Souza, 2014),
- Coefficients $(0, 1, \ldots, 14)$ of continuous wavelet transform with Ricker wavelet for $widths = (2, 5, 10, 20)$ (Mallat, 1999),
- All the coefficients (real and imaginary part, angle and absolute) of the fast Fourier transformation (FFT),
- Statistics of the absolute FFT (mean, variance, skew and kurtosis),
- Binned entropy of the power spectral density with the Welch method (Welch, 1967) ,
- Friedrich polynomial coefficients (Friedrich et al., 2000) for order of 3,
- Value of the partial autocorrelation function (Box, Jenkins, Reinsel, & Ljung, 2015) for $lag = (1, 2, \ldots, 10)$,
- Permutation entropy (Bandt & Pompe, 2002) with $dimension = (3, 4, \ldots, 7)$,
- Sample entropy (Richman & Moorman, 2000),
- Time reversal asymmetry statistic (Fulcher & Jones, 2014) with $lag = (1, 2, 3)$.

(The values chosen for the parameters of the aforementioned features are the commonly used values since tuning their values would require domain knowledge, defeating the purpose of a domain-agnostic framework).
After all of the features are extracted, to limit the number of irrelevant features, the FRESH algorithm (Christ, Kempa-Liehr, & Feindt, 2016) is deployed. It first performs the Kolmogorov-Smirnov test (Massey Jr, 1951) independently for every feature and calculates the p-value. Then, the FRESH algorithm utilizes the Benjamini-Yekutieli (Benjamini & Yekutieli, 2001) procedure under correction for dependent hypotheses to decide which null hypothesis H_0 to reject. Only the features for which the H_0 is rejected are kept. Finally, a Pearson correlation analysis is performed to remove features that are correlated with a value greater than 0.6, as this would indicate that they are (weakly) linearly correlated. Correlated features will get overweighted during the training, thus creating biased models whose results and generalizability can be compromised.

3.3.2. Expert-knowledge-based FE

Contrary to the first FE method, where a plethora of well-known features for sequential data are automatically extracted and filtered, for the expert-knowledge-based FE, as the name would suggest, a closer look at the data is required. After examining a normalized sequence for both a known good and a known bad coil (Figure 1), it becomes evident that the discrepancy between the two different classes is apparent in the time domain. Thus, the features that will be extracted are going to be limited to the time domain, meaning that no transformations to the data will be performed. The prominent characteristics of coils with a compromised surface quality are that they overshoot the upper and/or lower bounds of the accepted temperature range, that they present drifts from the goal temperature and, more importantly, they present abrupt peaks of high amplitude.
Based on the above observations, we choose to extract the features presented on Table 1. On the left column, the name of the feature is presented, while on the right are the values of the parameters that are used for their calculation. It is worth noting that the values of 0.05 and -0.05 were chosen for the threshold of the $count\_above$, $count\_below$ and $number\_crossing\_m$ features since the acceptable temperature range for the chosen steel grade is $\pm 5\%$.
The $number\_high\_peaks$ feature was engineered by the authors for this specific use case. The appearance of high-amplitude peaks is deemed detrimental to the classification of the coils, so a new feature is introduced to identify the peaks that have a standard deviation larger than 2 and return their count. The pseudo-code for the implemented feature can be found in Appendix.

3.4. Synthetic Labelling

Classification tasks are, by nature, handled by supervised algorithms. Supervised algorithms depend on labelled data to learn the decision boundary of the multidimensional manifold upon which the data points lie. To that end, the Tata Steel experts provided a set of 14 sequences of the steel grade under

Table 1. Features extracted for expert-knowledge-based FE.

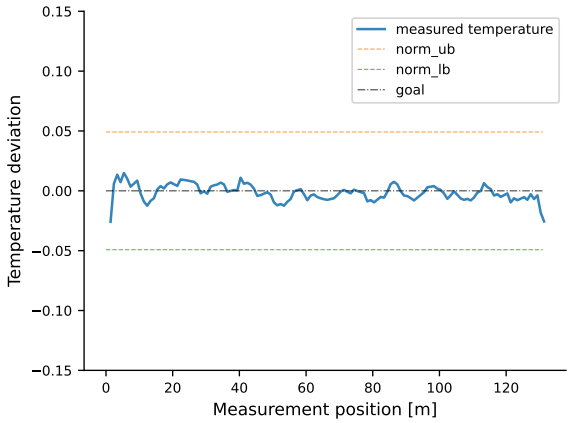
Feature	Parameters
abs_energy	-
absolute_maximum	-
absolute_sum_of_changes	-
cid_ce	normalize = False
count_above	t = 0.05
count_below	t = -0.05
skewness	-
longest_strike_above_mean	-
longest_strike_belowe_mean	-
maximum	-
mean	-
mean_abs_change	-
mean_change	-
minimum	-
number_crossing_m	m = 0.05 m = -0.05
standard_deviation	-
number_high_peaks*	n = 2 n = 5 n = 10

consideration that were identified to have low surface quality (one of which is shown in Figure 1b). Since the amount of labelled data is deemed inadequate to train a classification algorithm, the need to populate them arises. Upon inspection, and due to its use in the clustering analysis (Section 3.1), the DTW similarity metric is utilised. An ideal coil's CT profile would be identical to the goal CT profile. Leaning on that idea, the DTW similarity of each coil to the goal CT is calculated using Eq. (2). 76 coils with the highest score (indicating the **highest** dissimilarity to the goal CT) combined with the 14 expert-annotated ones comprise the bad coils labelled dataset. The 90 coils with the lowest DTW score form the good coils dataset.

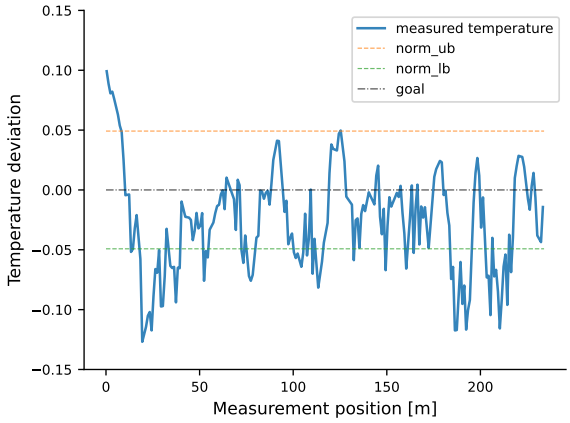
To guide the learning of the decision boundary, aside from providing examples of the extreme cases of both classes, we devised a labelled dataset of an extra 20 coils with intermediate DTW scores, half of which are used for training and half for testing. (this dataset will be referred to as manually annotated). This aims to not only provide information about the more ambiguous cases during training but also to provide a challenging test dataset that will assist in the evaluation of the performance of the classification algorithm. Figure 2 shows two of these coils. In conclusion, the final training dataset is constructed by performing an 80/20 % random split on the initial 180 coils and then adding half of the manually annotated dataset. The test dataset consists of the remaining data.

3.5. Neural Network classifier

A simple multilayer perceptron (MLP) is employed for the classification task. MLPs are fully connected feedforward NNs with non-linear activation functions. For the architecture of the model, typical design guidelines were followed. It consists of:



(a)



(b)

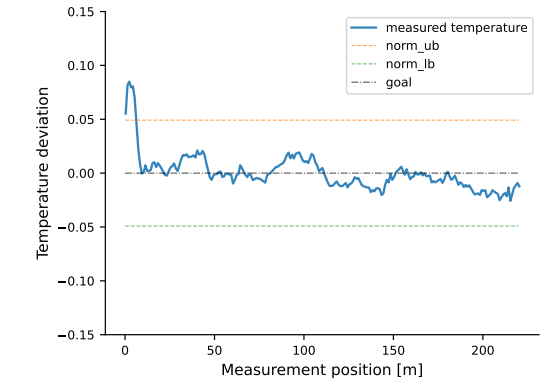
Figure 1. Normalized CT profile examples of a (a) good and a (b) bad coil

- **Input Layer:** where each feature is used as input for one input node,
- **Hidden Layer:** with $size = 64$, $relu$ activation function and to avoid overfitting, a dropout layer (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) with a dropout probability equal to 0.5,
- **Output Layer:** where, according to standard binary classification practice, it has $size = 2$ and a $softmax$ activation function, which will output the membership probability of each sample to each class.

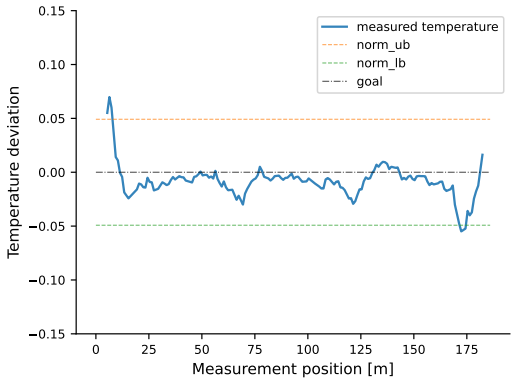
The simple and shallow architecture of the NN was chosen not only due to its decreased computational cost but also to avoid the tedious tuning and training of deep architectures.

4. Results

As previously discussed, the clustering analysis is performed on the raw data, while the classification is performed on the features extracted from the normalized sequences as described



(a)



(b)

Figure 2. Normalized CT profile examples of ambiguous coils. The spike in the beginning is observed on most coils, so it's not an indication of bad surface quality. Thus, (a) is labelled as good and (b) as bad (due to the peak at the end)

in Section 3.2. All of the code is written in Python, and the NN model was developed using Pytorch. Prior to the training of the NN, the training data are z-normalized. To avoid data leakage, the test data are z-normalized separately from the training data, utilizing the calculated scaling parameters of the training data. Models are trained for 200 epochs or until there is no improvement in the test accuracy. After the models have converged, they are tasked with classifying the entire dataset with all of the coils produced for the steel grade under consideration. The entire dataset follows the same FE procedure as the training set and is z-normalized with the pre-trained scaling parameters. Coils with a membership probability of less than 0.6 to either class are manually incorporated into the bad coils class to enhance our confidence in the models' predictions. Since the good/bad coil classification is the first step towards applying a PHM framework, we can tolerate false negatives, but we would like to avoid false positives. Since the good coils are of no interest to the analysis, a more inclusive bad coil class is preferred. First, the results of the clustering analysis on the raw data will be showcased, followed by the classification results with the introduced FE techniques.

4.1. K-means wth DTW

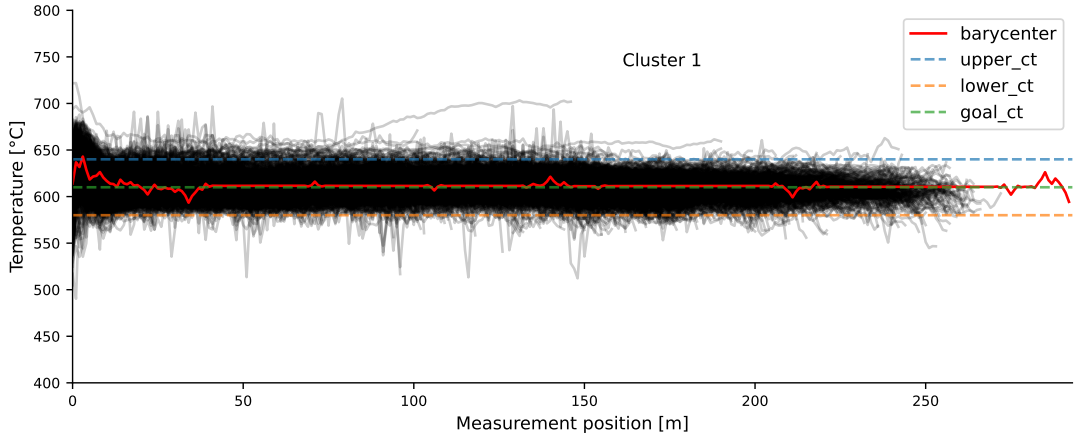
With the K-means algorithm, the number of cluster centres needs to be chosen a priori. To ensure the best fit, the elbow method is applied utilizing the silhouette coefficient (Rousseeuw, 1987). The results can be found in Table 2. As expected, the optimal number of clusters is two, confirming the prior assumption that the coils are split into good and bad ones. Figure 3 shows the results from the clustering and the calculated barycenters from the DBA algorithm. It becomes apparent that the majority of the coils in cluster 1 stay inside or close to the temperature boundaries, while bigger deviations are observed in cluster 2. This leads to the conclusion that the first cluster represents the good coils while the second cluster, the bad ones. However, the clustering is far from perfect since coils with high deviations and rapid fluctuations can be observed in the first (good) cluster. Given that the clustering analysis is the first exploratory step towards separating the data in hand, the results are satisfactory in that the expected underlying pattern of the data is actually observed. The high number of miss-clustered coils and the lack of soft-assignment capabilities means that it cannot be used as an end-to-end way to separate the data.

Table 2. Results of elbow method for DTW k-means.

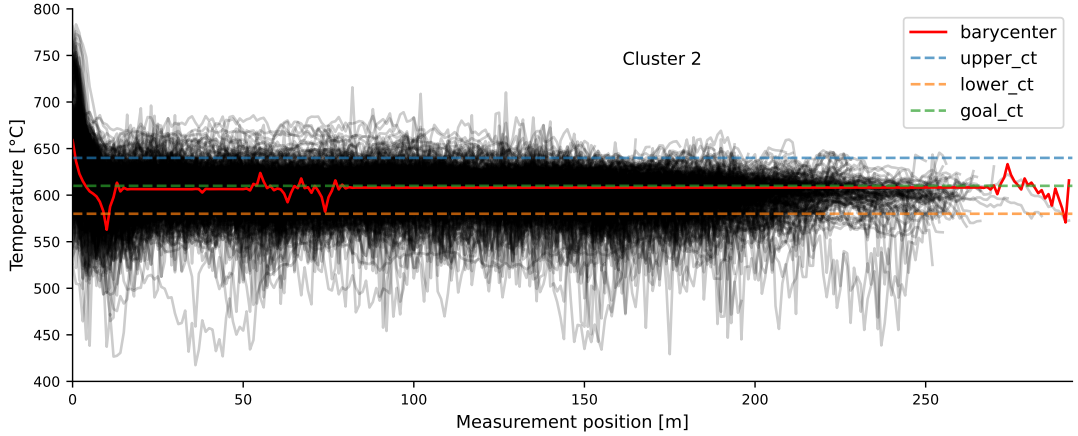
Clusters	Silhouette Score
2	**0.2556**
3	0.1938
4	0.1844
5	0.1947
6	0.1454

4.2. Classification with domain-agnostic features

After following the procedure of the domain-agnostic FE and filtering explained in Section 3.3.1, 111 features are left. The mean achieved accuracy of the NN is 0.8274 over the test data with 0.0180 standard deviation for 10 runs. The results can be seen in Figure 4. It can be observed that while the coils assigned in the bad class show a greater overall deviation from the goal CT, a lot of misclassified coils can be observed in the good class with highly fluctuating temperatures. This performance was to be expected, considering the rather low classification accuracy. To comprehend the low performance of the classification model, a principal component analysis (PCA) was performed on the extracted features with the goal of projecting the samples in a two-dimensional space. The calculated decision boundary is also drawn to enhance this visualisation's information. In order to achieve acceptable classification performance, the different classes need to present minimal overlap on the PCA space so that the classifier can find a way to separate them. This visualization can be seen in Figure 4c. A high overlap between the good and bad coils can be seen, meaning that no possible decision boundary can correctly separate the two classes, regardless of the choice of the model.

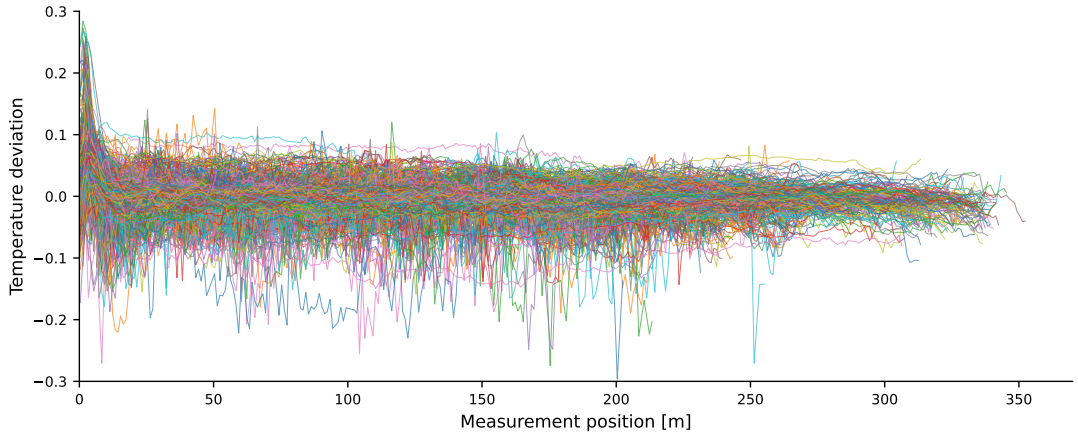


(a) Good coils cluster

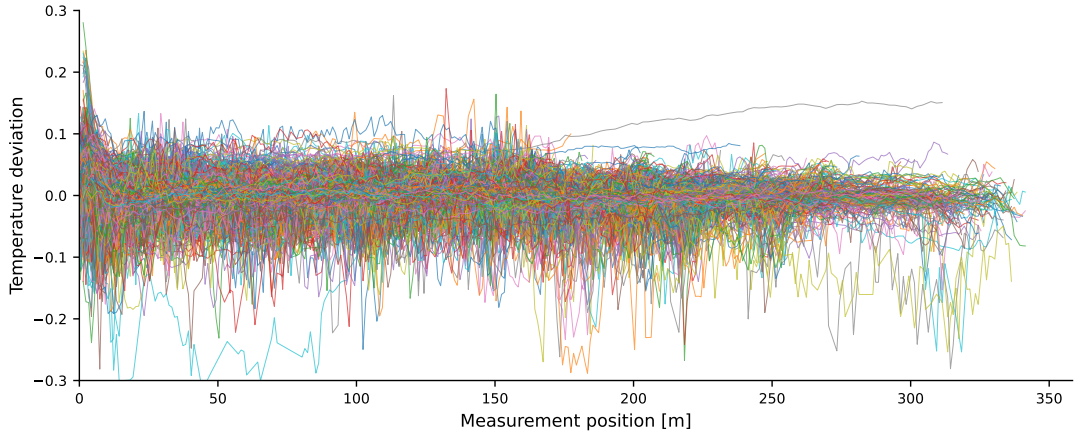


(b) Bad coils cluster

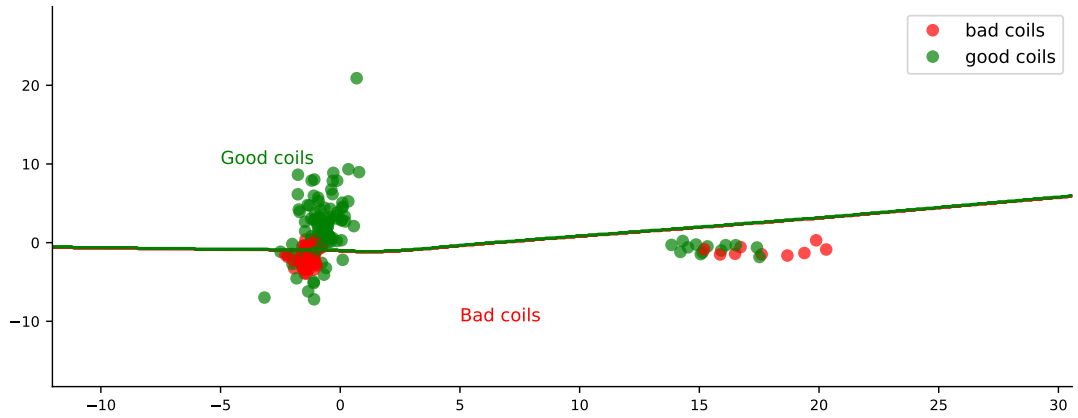
Figure 3. Clustering results with DTW K-means



(a) Good coils class

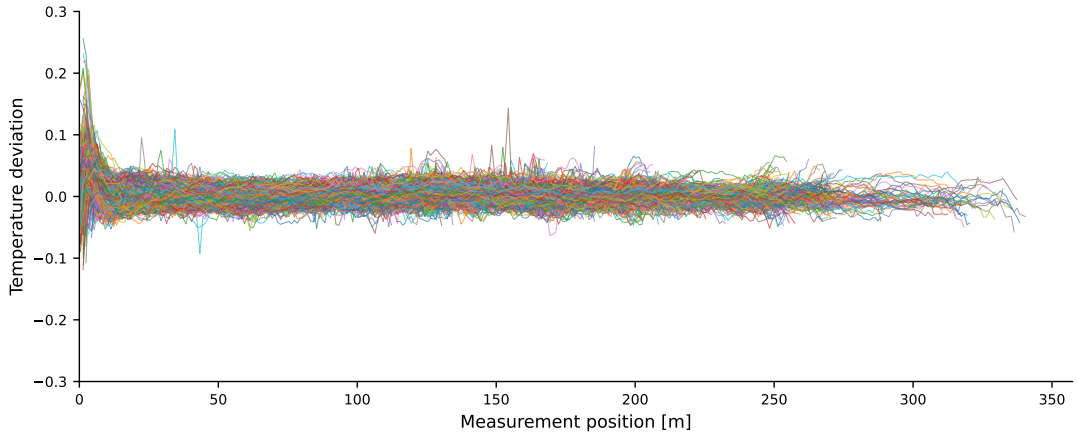


(b) Bad coils class

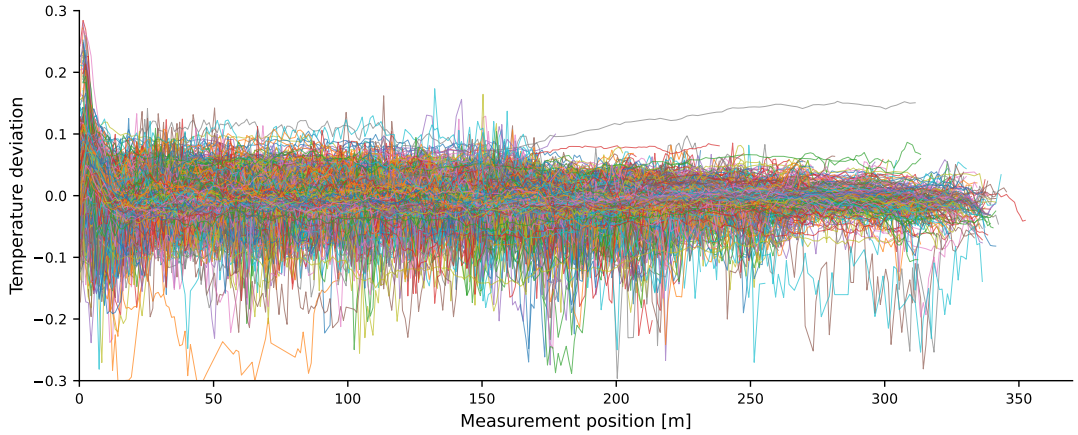


(c) 2-D PCA projection of the train and test samples

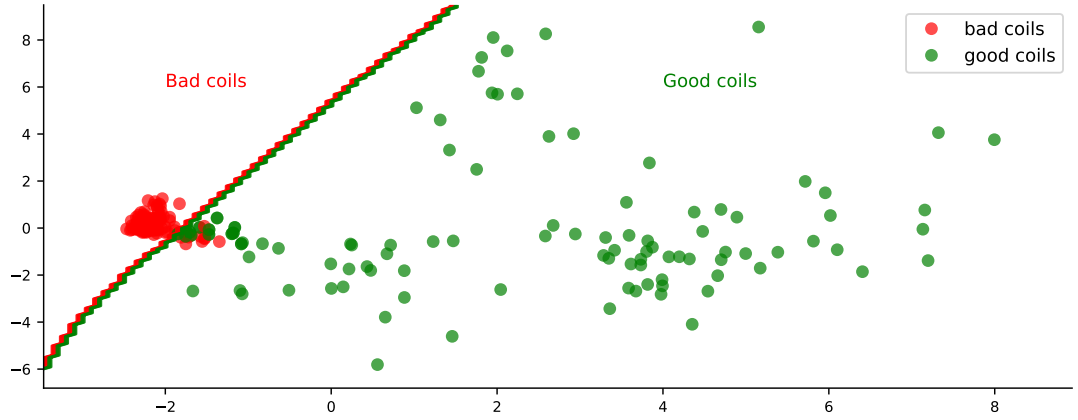
Figure 4. Classification results with domain-agnostic FE



(a) Good coils class



(b) Bad coils class



(c) 2-D PCA projection of the train and test samples

Figure 5. Classification results with expert-knowledge-based FE

4.3. Classification with expert-knowledge-based features

Following the FE method described in Section 3.3.2, 20 features are extracted. The same NN classifier architecture is used with the only change in the number of input nodes, which is altered to 20 to match the number of extracted features. The mean achieved accuracy is 0.9636 over the test data with 0.0075 standard deviation for 10 runs. The results can be seen in Figure 5. It becomes pretty apparent that the classification of the coils is superior to all of the previously presented methods. The healthy coils present minimal deviation from the goal temperature, with only a few coils that have a single abrupt temperature fluctuation in their CT, attesting to the high accuracy of the classifier. This means that there is a very limited number of false positive coils, which is highly important, as discussed at the beginning of the current section. The same visualization procedure is followed as before and presented in Figure 5c. It can be seen that there is a clear separation between the two classes and that the decision boundary lies optimally between them. This clear separation of the two classes explains the high performance of the rather simple classification model.

5. Discussion

The presented results pave the way for an important discussion when it comes to handling real-world complex and big data. After the clustering analysis was performed, the two expected different classes of coils could be identified, that is, the good and the bad class (referring to the CT profile and, in turn, the surface quality). However informative the clustering was in providing insight into the dataset, its performance was far from acceptable, with a lot of misclassified (or rather miss-clustered) coils. This is attributed to the fact that the K-means with DTW distance metric is clustering coils strictly by comparing their shape to each other. No information regarding the acceptable temperature range, the goal CT or what good and bad coils are, is included. Adding to that, the k-means algorithm does not provide a way to soft-assign clusters to data points. Naturally, the next step would be to increase the classification's performance will achieving soft-classification capabilities. The most obvious idea is to create a representation of the data to train a soft ML classifier. Due to the increasing popularity of DL FE methods, a researcher would most probably invest their time in developing complex and computationally heavy models. These models' task would be to try and learn on their own latent representations of the data that would effectively separate the different classes. With this study, we would like to emphasize that traditional FE can be as (if not more) effective for some datasets while reducing the complexity, the computational load, and the overall time invested in developing the FE method.

This is not to say that traditional FE can be applied universally, without effort. This is the main takeaway from comparing an automated traditional FE method that is domain-agnostic with features that are specifically picked or engineered for the application. For the domain-agnostic FE, a plethora of famous and commonly used features for sequential data were automatically extracted and filtered utilizing hypothesis tests and correlation analysis. However, the resulting features fail to capture the distinctive features of the data. This becomes evident by the high overlap of the two classes presented in Figure 4c, and is the culprit of the wrong classification of the data. Spending the effort of manually labelling a small fraction of coils and choosing the right features to represent the data, successfully separates them and achieves the required classification performance.

The next step for this framework is to verify that it works universally for multiple steel grades, with minor or even no modifications at all. After generating the healthy/damaged coils dataset, the process parameters that lead to the damaged state are intended to be identified. The end goal is to apply a PHM framework that will be able to predict quality deterioration and provide alternative parameter settings to mitigate the damage to the surface quality of the produced steel strips.

6. Conclusions

In this study, a real-world data set of manufactured steel strips raises the importance and effectiveness of traditional FE, but only if done appropriately, as described in Section 3.3.2 and paired with manually annotated samples. Automated FE techniques are deemed ineffective; thus, the extracted features must be chosen carefully. This is achieved by keeping in mind that they should capture the characteristics that associate them with their corresponding class. The authors are by no means undermining the importance of deep learning FE methods. Their increasing popularity mainly stems from their successful application in extracting latent representations of big data. They would instead highlight that for some datasets, the effort needed to develop them is unjustified; that is when a correctly defined traditional FE method can solve the task.

References

Amari, S.-I. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on computers*, *100*(11), 1197–1206.

Bandt, C., & Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical review letters*, *88*(17), 174102.

Batista, G. E., Keogh, E. J., Tataw, O. M., & De Souza, V. M. (2014). Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, *28*, 634–669.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*.

John Wiley & Sons.

Caesarendra, W., & Tjahjowidodo, T. (2017). A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing. *Machines*, *5*(4).

Chen, R., & Yuen, W. (2001). Oxide-scale structures formed on commercial hot-rolled steel strip and their formation mechanisms. *Oxidation of metals*, *56*(1), 89–118.

Choi, S. W., Lee, C., & Lee, e. a. (2005). Fault detection and identification of nonlinear processes based on kernel pca. *Chemometrics and intelligent laboratory systems*, *75*(1), 55–67.

Christ, M., Kempa-Liehr, A. W., & Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*.

Cuturi, M., & Blondel, M. (2018). *Soft-dtw: a differentiable loss function for time-series.*

Deng, G., Zhu, H., Tieu, A. K., Su, L., Reid, M., Zhang, L., ... others (2017). Theoretical and experimental investigation of thermal and oxidation behaviours of a high speed steel work roll during hot rolling. *International Journal of Mechanical Sciences*, *131*, 811–826.

Friedrich, R., Siegert, S., Peinke, J., Siefert, M., Lindemann, M., Raethjen, J., ... others (2000). Extracting model equations from experimental data. *Physics Letters A*, *271*(3), 217–222.

Fulcher, B. D., & Jones, N. S. (2014). Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 3026–3037.

Gush, T., Bukhari, S. B. A., Haider, R., Admasie, S., Oh, Y.-S., Cho, G.-J., & Kim, C.-H. (2018). Fault detection and location in a microgrid using mathematical morphology and recursive least square methods. *International Journal of Electrical Power & Energy Systems*, *102*, 324–331.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, *20*(7), 1483–1510.

Jiang, W., Hong, Y., Zhou, B., He, X., & Cheng, C. (2019). A gan-based anomaly detection approach for imbalanced industrial time series. *IEEE Access*, *7*, 143608–143619.

Kankar, P. K., Sharma, S. C., & Harsha, S. P. (2011). Fault diagnosis of ball bearings using continuous wavelet transform. *Applied Soft Computing*, *11*(2), 2300–2312.

Liu, H., Zhou, J., Xu, Y., Zheng, Y., Peng, X., & Jiang, W. (2018). Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks. *Neurocomputing*, *315*, 412–424.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Min, K., Kim, K., Kim, S. K., & Lee, D.-J. (2012). Effects of oxide layers on surface defects during hot rolling processes. *Metals and Materials International*, *18*, 341–348.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, *44*(3), 678–693.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology-heart and circulatory physiology*, *278*(6), H2039–H2049.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Sakoe, H. (1971). Dynamic-programming approach to continuous speech recognition. In *1971 proc. the international congress of acoustics, budapest.*

Schreiber, T., & Schmitz, A. (1997). Discrimination power of measures for nonlinearity in a time series. *Physical Review E*, *55*(5), 5443.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Wang, Y., Pan, Z., Yuan, X., Yang, C., & Gui, W. (2020). A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network. *ISA transactions*, *96*, 457–467.

Wang, Y., Yang, H., Yuan, X., Shardt, Y. A., Yang, C., & Gui, W. (2020). Deep learning for fault-relevant feature extraction and fault classification with stacked supervised auto-encoder. *Journal of Process Control*, *92*, 79–89.

Wang, Z., McConnell, S., Balog, R. S., & Johnson, J. (2014). Arc fault signal detection - fourier transformation vs. wavelet decomposition techniques using synthesized data. In *2014 ieee 40th photovoltaic specialist conference (pvsc)* (p. 3239-3244).

Wang, Z., Wang, J., & Chen, S. (2020). Fault location of strip steel surface quality defects on hot-rolling production line based on information fusion of historical cases and process data. *IEEE Access*, *8*, 171240-171251.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, *15*(2), 70–73.

Wu, H., & Zhao, J. (2018). Deep convolutional neural network model based chemical process fault diagnosis. *Comput-*

ers & chemical engineering, *115*, 185–197.

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X., & Ding, N. (2022). Gan-based anomaly detection: A review. *Neurocomputing*, *493*, 497–535.

Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., & Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Annals of biomedical engineering*, *41*, 349–365.

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237.

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering System Safety*, *218*, 108-119.

APPENDIX

Algorithm 1 Pseudocode of number_high_peaks feature

Inputs:
x (list): the input sequence
n (int): the support of the peak (a peak of support n is defined as a subsequence of x where a value occurs, which is bigger than its n neighbours to the left and to the right)
std_t (int): the number of standard deviations that the peak's value needs to surpass
Procedure:
$x\_reduced = x[n : -n]$
$res = None$
for $(c = 0; c < n + 1; c + +)$ **do**
  $result\_first = x\_reduced > numpy.roll(x, c)[n : -n]$
  if $res = None$ **then**
    $res = result\_first$
  else
    $res \mathrel{+}= result\_first$
  end if
  $res \mathrel{+}= x\_reduced > numpy.roll(x, c)[n : -n]$
end for
$idx\_peaks = np.where(res)[0] + n$
$h_peaks = 0$
for $idx : idx\_peaks$ **do**
  if $|x[idx]| > mean(x) + std\_t * std(x)$ **then**
    $h\_peaks \pm 1$
  end if
end for
Output:
h_peaks (int): the amount of peaks of support n with maximum value higher than std_t times the standard deviation of x

Towards a Hybrid Framework for Prognostics with Limited Run-to-Failure Data

Luc S. Keizers[1,2], Richard Loendersloot[1], Tiedo Tinga[1,2]

[1] *University of Twente, Enschede, 7522NB, the Netherlands*
l.s.keizers@utwente.nl,
r.loendersloot@utwente.nl,
t.tinga@utwente.nl

[2] *Netherlands Defence Academy, Den Helder, 1781AC, the Netherlands*

ABSTRACT

The introduction of cyber-physical systems with increased availability of sensor data creates a lot of research interest in prognostic algorithms for predictive maintenance. Although a lot of algorithms are successfully applied to benchmark case studies based on simulated data and experimental set-ups, deployment in industry lags behind. From a comparison between three benchmark case studies with two real-world case studies based on prognostic metrics (monotonicity, prognosability and trendability), two main issues are observed: 1) the lack of run-to-failures and 2) low prognostic metrics due to a low signal-to-noise ratio of degradation trends, as a result of unexplained physical phenomena. To make prognostics feasible, a hybrid framework is proposed that focuses on improving system knowledge. The framework consists of a quantitative diagnostic assessments, guided by (modular) system models in which damage is induced. This quantitative damage assessment provides input for prognostics based on Bayesian filtering, enabling prognostics for assets in varying operational conditions. Implementation and validation of the framework requires investments, but modularity within the framework can accelerate development for new systems.

1. INTRODUCTION

During the fourth industrial revolution, cyber-physical systems are being introduced where sensor data communicates between machinery and with operators (Pinciroli et al., 2023). This sensor data can be used to find characteristics of failures within the systems which can be used to develop models that predict the remaining useful life (RUL) (Yan et al., 2017), giving new opportunities for implementation of *predictive maintenance*. When failures can be predicted, catastrophic accidents are prevented, unexpected downtime can be reduced, components are used until the end of their actual lifetime and maintenance logistics can be optimized (Fernandes et al., 2022).

The recent increasing interest in *predictive maintenance* is clearly visible by observing the amount of published scientific papers in this field. Only the number of review papers is already growing significantly, as the number of counts in the Scopus database on article titles with (*survey* or *review*) and (*predictive maintenance*) grew from a total of 20 published documents up to 2020 to a total of 84 published documents up to 2023.

The increasing number of sensors and data availability, specifically increase the interest in *data-driven* prognostic approaches (Pinciroli et al., 2023). This type of approach requires sufficient historical run-to-failure data. However, in safety-critical systems (Chao et al., 2021) or when availability of assets is more important than costs (Tinga et al., 2021), failures are sparse and as a consequence the required historical run-to-failure data are rarely being collected. Also for new types of machinery, no historical data are available (Calabrese et al., 2021). If historical data are available, they are often unlabeled and unorganized, lacking context such as operating conditions and maintenance recordings (Calabrese et al., 2021; Lukens et al., 2022).

In contrast to data-driven approaches, physics-based approaches considering Physics-of-Failure (PoF) models have less strict data requirements. They provide a relation between usage and degradation rates (Tinga, 2013b). This yields benefits compared to purely data-driven approaches, specifically when failures are rare and when future operating conditions (and consequently degradation rates) are different from historical operating conditions (Tiddens et al., 2023). However, these models are expensive to develop and are component or system specific (Elattar et al., 2016). Also, the relation between usage and degradation rates should be known and must not be too

complex.

To overcome issues with purely data-driven or purely physics-based prognostics, combinations of them (i.e. hybrid approaches) are often proposed as a solution (Elattar et al., 2016; Guo et al., 2020). Still, many hybrid methods use data-driven models to estimate the degradation behavior (Pugalenthi et al., 2021; Borutzky, 2020) or only use physics to improve input parameters of a data-driven prognostic algorithm (Gálvez et al., 2021; Chao et al., 2022). This yields improved results compared to purely data-driven approaches, but the relation between usage and degradation rates are generally not considered, still limiting applicability in cases of rare historical failures and varying operating conditions.

Fernandes et al. (2022) described that in many studies, real-world challenges are often overlooked and additional research to address these real-world challenges is needed. This suggests a gap between science and industry. This suggestion is strengthened by a survey in 280 companies in Belgium, Germany and the Netherlands, showing that only 11% are actually implementing predictive maintenance techniques in 2017 (Mulders & Haarman, 2017). Although a similar survey showed that this number increased to 17% in 2023, mainly attributed to original equipment manufacturers (OEMs) and sectors with large numbers of the same assets (van der Velde et al., 2023), a large majority of developed methods is still only being applied to experimental or simulated data sets (Ferreira & Gonçalves, 2022) such as C-MAPSS (Saxena & Goebel, 2008).

In this paper, issues with implementation of prognostics in two real-world cases are pinpointed. The case studies are performed by two organizations within the Netherlands who try to implement predictive maintenance for military assets. The case studies are challenging, as availability of these assets is more important than costs (Tinga et al., 2021), fleets are relatively small and the assets operate in varying operational and environmental conditions. Prognostic metrics defined by Coble (2010) (monotonicity, prognosability and trendability) are calculated for the two real-world case studies, and also for three well-known benchmark cases to compare the potential for application of prognostic algorithms. Based on observed issues in the real-world cases, a possible solution is proposed in the form of a hybrid framework.

The remainder of the paper is organized as follows. Section 2 starts with calculating prognostic metrics of features from three benchmark cases: the Virkler crack growth data set, the NASA milling data set and the C-MAPSS data set. Then, metrics are calculated for features from the two real-world case studies, concerning condition monitoring of Apache helicopter engines and a naval main diesel engine respectively. Following from the issues observed in the case studies, section 3 proposes a hybrid framework for prognostics. Lastly, section 4 discusses the results and concludes the paper.

2. Prognostic Potential of Case Studies

2.1. Prognostic Metrics

Coble (2010) developed three metrics to assess the suitability of features as input for a (data-driven) prognostic algorithm. The suitability is evaluated based on monotonicity (M), prognosability (P) and trendability (T). The range of the scores is from 0 (unsuitable) to 1 (perfectly suitable). The weighted sum of the three metrics gives the prognostic score, and features with the highest score are most suitable to be used as the input for a prognostic algorithm. So, data sets from which features with high scores can be derived, have high potential for prognostics.

The first metric is monotonicity, assessing the extent to which run-to-failure trajectories are purely increasing or decreasing. It is calculated as follows:

$$M = \text{mean}\left(\left|\frac{N^+ - N^-}{n-1}\right|\right) \tag{1}$$

with N^+ the number of increments in the run-to-failure trajectory of the feature (i.e. $n_{i+1} - n_i > 0$), N^- the number of decrements in the trajectory (i.e. $n_{i+1} - n_i < 0$) and n the number of data points in the trajectory. The absolute mean monotonicity of all considered run-to-failure trends yields the final monotonicity.

Prognosability estimates how similar the start values and the values at failure are for the features when comparing different run-to-failure trajectories. It is calculated as follows:

$$P = \exp\left(-\frac{\text{std}(\mathbf{f}_{end})}{\text{mean}\left(|\mathbf{f}_{end} - \mathbf{f}_0|\right)}\right) \tag{2}$$

with $\mathbf{f}_{end}$ a vector with all values of the features at failures and $\mathbf{f}_0$ a vector with all values of the features at the start of the run-to-failure trajectories. Std refers to the standard deviation.

Trendability describes similarity between the shapes of run-to-failure trajectories. It is calculated as follows:

$$T = \min\left(\left|\boldsymbol{\rho}_{ij}\right|\right) \tag{3}$$

with $\boldsymbol{\rho}$ a vector with the correlation coefficients between each run-fo-failure trajectory i and j of the feature. For trajectories with different lengths, linear interpolation is applied such that the lengths of the correlated trajectories match.

The final score S is calculated by:

$$S = W_m \cdot M + W_p \cdot P + W_t \cdot T \tag{4}$$

with W_M, W_P and W_T the weight factors for monotonicity, prognosability and trendability respectively. In many applications they can be set identically, but in some applications some metrics may be less relevant (Coble, 2010). In the case studies discussed in the next subsection, the weight factors are

all set to $\frac{1}{3}$, yielding prognostic scores ranging from 0 to 1.

2.2. Benchmark Data Sets

2.2.1. Virkler Crack Growth

The first benchmark data set considered is the Virkler crack growth data set. Virkler et al. (1979) performed 68 run-to-failure fatigue tests of 2024-T3 aluminium and measured the crack length directly. This yields the run-to-failure trajectories as shown in Fig. 1.

The data set contains crack lengths at specific numbers of stress cycles. As crack lengths already provide a direct indicator of the damage severity, no additional features need to be calculated and the prognostic metrics are directly calculated on the crack length measurements. As the crack length always increases monotonically, the monotonicity is 1. All trajectories have the same end value (50mm) and starting value (9mm), yielding a prognosability of 1. As all trajectories are monotonically increasing, all trajectories have a perfect positive correlation, yielding a trendability of 1. Consequently, the total prognostic score is 1.

A direct measurement of degradation can be considered as a perfect prognostic metric, as the nature of degradation (an irreversible process) makes it monotonic, a proper threshold can be defined based on system knowledge and the monotonicity yields also perfect trendability. As the underlying model is well understood, the Virkler data set is perfectly suitable for prognostic methods based on Bayesian updating (Sun et al., 2014; Baral et al., 2023), but also data-driven methods are well applicable (Eker & Jennions, 2012).

2.2.2. Milling Tool Wear

The Milling Data Set (Agogino & Goebel, 2007) contains data collected from an experimental setup for tool wear estimation. There are two operational settings for the Depth of Cut (DOC), feed rate and material. Two experiments are performed for each combination of operational settings, yielding a total of 16 experiments. Each experiment consists of a number of runs, lasting 72s. After each run, the actual tool wear (i.e. *VB*: the measured distance from the cutting edge of the tool to the end of abrasive wear on the flank (Agogino & Goebel, 2007)) is measured with a microscope. Experiments were terminated at a certain (not further specified) tool wear limit (and some beyond). As measuring tool wear with a microscope after each operation is not feasible in practical applications, measurements are also collected from sensors that can provide an online estimation of tool wear. A total of six sensor are installed which measure at a sampling rate of 250Hz: AC and DC motor current sensors of the spindle, and vibration and acoustic emission sensors at both the spindle and the table.

The data set is shown in Fig. 2. Fig. 2a shows the offline tool wear depth (*VB*) measurements. As most experiments run until $\approx 0.50 - 0.80$mm as shown by the gray band in the figure, it is expected that the tool wear limit is around this band. The gaps in the trajectories in Fig. 2a are due to missing measurements in between some of the runs. Note that only 14 out of the 16 experiments are displayed: for one of the experiments, only one run is available, making it unsuitable to calculate prognostic metrics. Both experiments in the corresponding

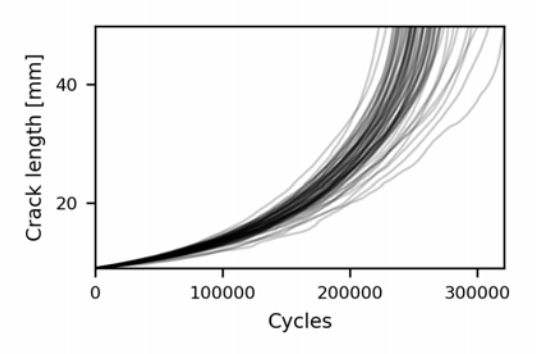
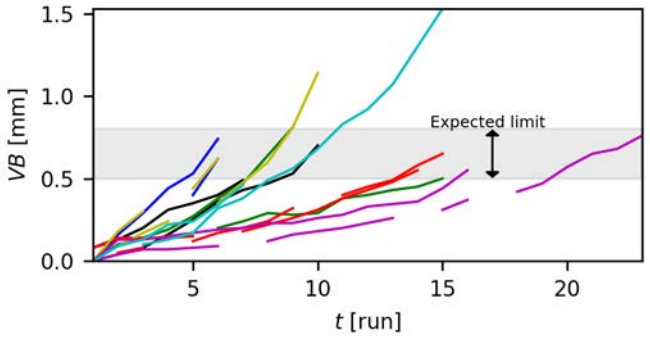
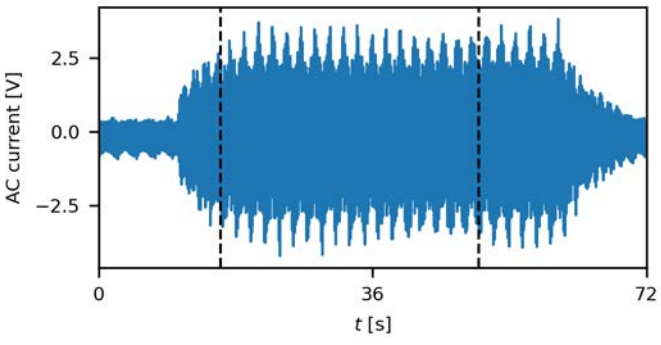


Figure 1. Virkler Crack Growth Data Set Virkler et al. (1979) (M = 1, P = 1, T = 1, S = 1)



(a) All tool wear trajectories in milling data set. Each color corresponds to a set of operational settings (M = 0.95, P = 0.66, T = 0.78, S = 0.80)



(b) Example (AC motor current) measurements for one run in milling data set

Figure 2. Visualization of the milling data set

operational conditions are removed from the data set, such that seven pairs of experiments remain.

The trajectories for the *VB* measurements yield a monotonicity M of 0.95, prognosability P of 0.66, trendability T of 0.78 and S of 0.80. Although perfect metrics of 1 are expected for direct degradation condition measurements, as explained in subsection 2.2.1, some non-monotonic behavior can be observed in Fig. 2a (e.g. the dark green line before $t = 10$), reducing M and T. As wear is irreversible, this non-monotonic behavior is likely to be due to measurement errors. P is affected by the fact that some experiments were performed beyond the tool wear limit. Still, the prognostic metrics are high, and can be improved by reducing measurement error and running experiments until a fixed failure threshold.

A challenge is to estimate tool wear from the real-time sensor data. Fig. 2b shows an example of data from the AC motor current sensor for one run of an experiment. No clear trend can be observed in the raw sensor data, so features need to be calculated. As the start and run of an experiment yield no stable signal (i.e. magnitude increase and decrease as seen in Fig. 2b), only the stable period (defined to be 16-50s, indicated by the vertical dashed bars in Fig. 2b) is considered for feature calculation. For each sensor and each run, the mean, standard deviation, maximum, minimum, absolute maximum, absolute minimum, root mean squared and sum of values are calculated.

The standard deviation from the AC motor current measurements is found to have the highest prognostic score and its trajectories are shown in Fig. 3. M is 0.74, P is 0.57 and T is 0.84, yielding a score of 0.72. The different operating conditions clearly yield different feature values, as only the start and end values of curves with the same operating condition have approximately the same start- and end values (i.e. the dark green, black, red and pink pairs). Prognostic scores can be further improved by compensating for operating conditions, by calculating the P and T for two experiments with the same

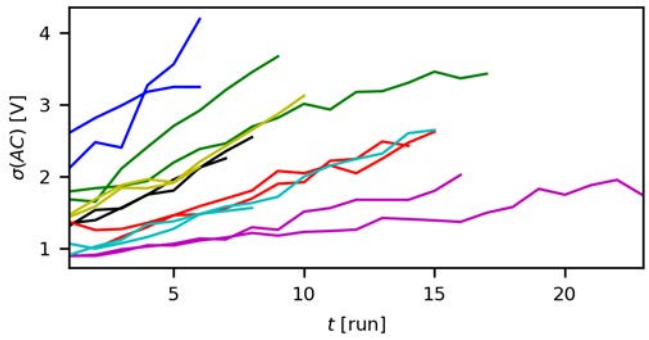


Figure 3. Feature with highest score from milling data set, color-coded by operating conditions (M = 0.74, P = 0.57, T = 0.84, S = 0.72)

operating conditions and taking the mean of the separately calculated P and T (note that M is unaffected). It is found that in this case, P increased to 0.78 and T increased to 0.86, yielding a final score of 0.79, which is almost the same score as for direct *VB* measurements.

This case shows that by calculating only a simple set of features, features with high prognostic potential can already be obtained. These characteristics make the data set applicable for prognostics. In literature, mainly quantitative diagnostic methods are applied, based on e.g. nearest neighbor-based approaches (Sheng & Zhu, 2020), Recurrent Neural Networks (Lu et al., 2022), Kernel Extreme Learning Machines (Zhou & Sun, 2020), particle filters (P. Wang & Gao, 2016) and Long Short-Term Memory Networks (Kumar et al., 2022). Subsequently, prognostics can be performed (J. Wang et al., 2015).

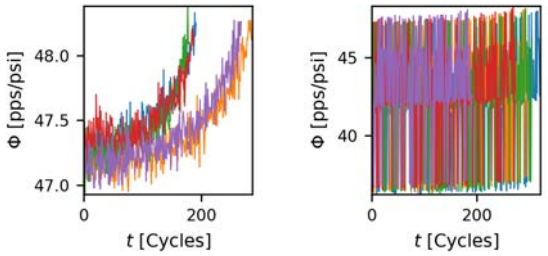
2.2.3. C-MAPSS

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) (Saxena & Goebel, 2008) is a very popular benchmark data set due to the inclusion of sensor noise, different operating conditions and multiple fault modes (Ramasso & Saxena, 2014). Already in 2014, Ramasso & Saxena (2014) published a review paper on 70 different prognostic methods utilizing the data set. The data set is still being used and more recently a new version of this data set (N-CMAPSS) has been released (Chao et al., 2021). The data is generated with a simulator built in Matlab and Simulink. The operational settings are defined by three parameters and 21 (virtual) sensors are available.

In this paper, the FD001 and FD004 train data sets are evaluated. The FD001 set contains 100 degradation trajectories in one operating condition and one fault mode (High Pressure Compressor (HPC) degradation). The FD004 set contains 248 degradation trajectories in six operating conditions and two degradation modes (HPC degradation and fan degradation). As an example, Fig. 4 shows raw sensor data from one of the sensors (Φ, a fuel flow ratio) for five run-to-failure trajectories of FD001 and FD004. In Fig. 4a the degradation trends can be clearly observed, which is more difficult in Fig. 4b due to the effect of changing operating conditions on the data.

To reveal the degradation trend for the FD004 data set, a K-Nearest Neighbors regressor is trained on the first 40 data points to learn the (nominal) relation between the three operational settings and measurements. The regressor is built using the $sklearn$ Python package and uses two neighbors. Fig. 5 shows that the residuals (i.e. difference between measured and expected measurements) reveal similar degradation trends as was observed for the FD001 data set.

Although the raw sensor data of FD001, or the residuals for FD004, already reveal a strong degradation trend, better prog-



(a) Five run-to-failure trajectories of Φ in constant operating conditions (FD001)

(b) Five run-to-failure trajectories of Φ in varying operating conditions (FD004)

Figure 4. Five run-to-failure trajectories of Φ in the C-MAPSS data set

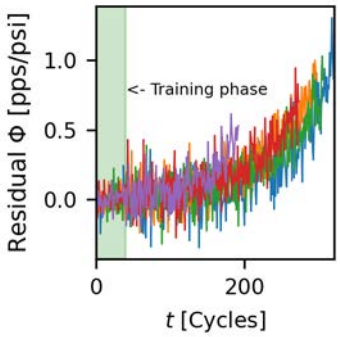
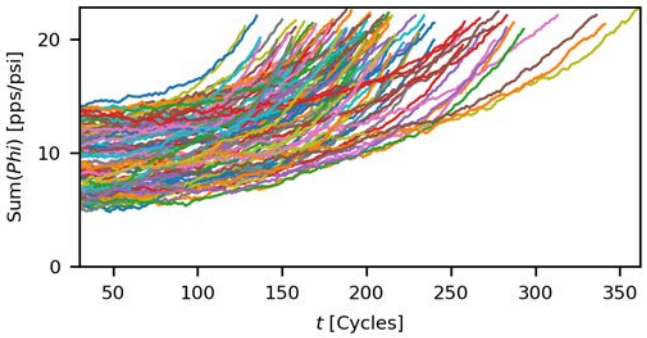
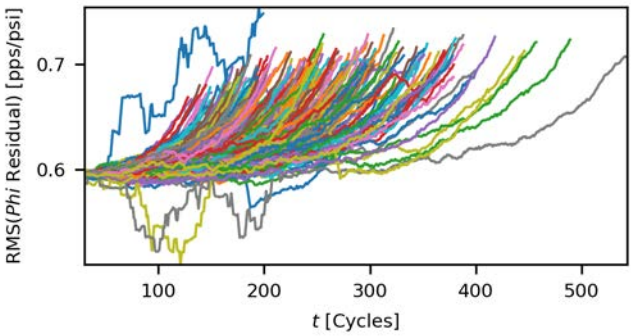


Figure 5. Five run-to-failure residual trajectories of Φ sensor in varying operating conditions (FD004)



(a) Top feature of C-MAPSS FD001 data set (directly derived from sensor data) (M = 0.69, P = 0.88, T = 0.94, S = 0.84)



(b) Top feature of C-MAPSS FD004 data set (derived from residuals) (M = 0.55, P = 0.80, T=0.06, S =0.51)

Figure 6. Best features on CMAPSS data set

nostic metrics can be found by extracting features. This is done for both the FD001 raw measurements and the FD004 residuals using the Python package *tsfresh* (Christ et al., 2018) for automatic feature extraction. It first creates a rolling window over the data set and for each window, features are calculated. The window size is a trade-off between noise reduction and response time to signal changes, and is set to 30 by trial-and-error. To reduce noise in the calculated features, the prognostic metrics are calculated for each 5th data point.

The best performing feature for the FD001 and FD004 data set are shown in 6. For FD001 the sum of Φ measurements in the rolling window obtained the highest metrics, as shown in Fig. 6a: M=0.69,P=0.94 and T=0.84, yielding a score of 0.84. For FD004 the root mean square of the Φ residuals gave the highest metrics, which are significantly lower (i.e. M=0.55,P=0.80 and T=0.06, yielding a score of 0.47) compared to FD001. This is mainly caused by some outlying trajectories (e.g. the blue trajectory on the left in Fig. 6b) due to faulty residual calculations (e.g. because not all operating conditions are observed in the training phase of the KN-regressor). It is found that the mean of the correlation coefficients of all trajectories is 0.97, but the fact that T is determined by the lowest correlation coefficient between all trajectories yields the low trendability. Therefore, higher scores can be obtained when removing outliers and improving residual generation.

To conclude, it is found to be straightforward to retrieve features with high prognostic metrics, mainly for the FD001 data set. For the FD004 data set it is more challenging due to the varying operating conditions. However, features could be extracted which show similar run-to-failures trajectories, although additional effort is required to remove outliers and improve prognostic metrics further. The general characteristics of the data set make it feasible for data-driven prognostics, and methods such as Nearest-Neighbors, Random Forests, Extreme Gradient Boosting and Multilayer Perceptrons (Alomari et al., 2023), Convolutional Neural Networks-based approaches, Long-Short-Term-Memory-based approaches (de Pater et al., 2022) and others are widely found in literature.

2.3. Real-world Case Studies

This subsection introduces two real-world case studies. Despite an extensive search, only two cases were found to have sufficient measurements and meta data available to calculate metrics of run-to-failure trajectories. Organizations often do

not have, cannot or do not want to disclose the data necessary for a proper analysis. The authors thank NLR and the Royal Netherlands Navy gratefully for making these case studies available for evaluation.

It should be noted that the case studies concern the most complex cases for prognostics: they concern monitoring of individual assets in varying operating conditions, where future operating conditions can be different from historical operating conditions. This corresponds to the highest ambition level an organization can have, with high requirements on data availability or system knowledge (Tiddens et al., 2023).

2.3.1. Apache ETF Monitoring

The power level of helicopter turboshaft engines decreases over the lifetime due to wear of seals, vanes and blades or due to faults in other components (Vos, 2019). Engine performance is measured by the Engine Torque Factor (ETF), which is the ratio between the actual engine power and the rated engine power. If the ETF drops below 0.85, or if the combination of the ETFs of both tail engines of an Apache helicopter drops below 0.90, the Apache is not allowed to be used. The Netherlands Aerospace Centre (NLR) developed an algorithm to calculate the ETF from in-flight parameters, rather than from time-consuming manual Max Power Checks (MPCs) (Vos, 2019).

Vos (2019) selected the turbine gas temperature (TGT) as health indicator, and fitted a polynomial model using data from the Health and Usage Monitoring System (HUMS) to translate the in-flight TGT to the TGT at the reference condition (which in turn allowed to calculate ETF). The considered operational parameters are gas inlet temperature, outside air temperature, pressure altitude, speed, and engine torque.

For the development of a prognostic algorithm, run-to-failure trajectories are required. Although the engine needs an overhaul when the ETF reaches 85%, the data set does not contain any trajectory running till this threshold. Therefore, to calculate the prognostic metrics of trajectories, periods between (documented) engine replacements are selected. This does not fully represent run-to-failure, but no better option is available.

The rolling mean of these ETF trajectories is calculated over five ETF measurements, yielding the trajectories in Fig. 7. The calculated metrics over these trajectories are: M=0.09, P=0.19 and T=0, yielding the extremely low score of 0.12. This can be expected from Fig. 7, as the data are covered in low-frequency noise which make the actual degradation trend barely visible. This low-frequent wobbling behavior is caused by physical phenomena not explained by the polynomial model (Vos, 2019), i.e. by confounding factors. This noise yields a low signal-to-noise ratio (i.e. degradation to other external influences), and therefore low prognostic metrics.

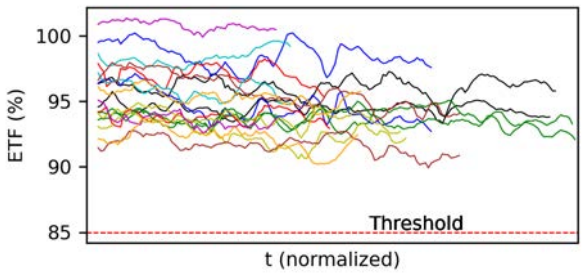


Figure 7. All ETF trajectories considered. Each color corresponds to the engines of a specific Apache. (M = 0.09, P = 0.19, T = 0, S = 0.12)

Real-time estimation of the ETF offers potential to replace expensive MPC by real-time ETF monitoring (i.e. additional inspections can be performed when the ETF drops below the threshold). However, the lack of actual run-to-failures and the low signal-to-noise ratio make the step towards prognostics extremely complicated.

2.3.2. Marine Diesel Engine Bearing Monitoring

The main diesel engine (MDE) of a naval vessel contains seven journal bearings that support the crankshaft. Failure of one of the bearings yields failure of the MDE and is therefore critical for availability of the vessel. Heek (2021) developed a monitoring method based on bearing temperature to detect failures timely. The underlying idea is that damage increases friction in the bearing and subsequently increases the bearing temperature.

Because the bearing temperature is also affected by operating conditions, Heek (2021) fitted data from the Integrated Platform Management System (IPMS) in a multiple linear regression model. This model estimates the bearing temperature in nominal conditions based on the RPM of the engine, RPM of the turbocharger, and the lube oil temperature at the outlet, which were found to have the highest predictive performance. Subsequently, the residual between the measured and the predicted bearing temperature is selected as a health indicator and is continuously monitored. Alarms can be raised when the measured temperature is higher than expected. The complete procedure of data selection and residual generation can be found in Heek (2021).

Heek (2021) evaluated three case studies, from which two concerned an actual failure. Cases 1 and 2, which concern failure cases, are visualized in Fig 8. The regression model is trained within the shaded areas (until $t = 993$ and $t = 1748$ respectively) and the residual monitoring is deployed after (note that case 2 runs longer than case 1). To evaluate prognostic metrics of these trajectories, every 10th data point is selected and the rolling mean over 10 data points is calculated. The calculated metrics are: M=0.06, P=0.92 and T=0.57,

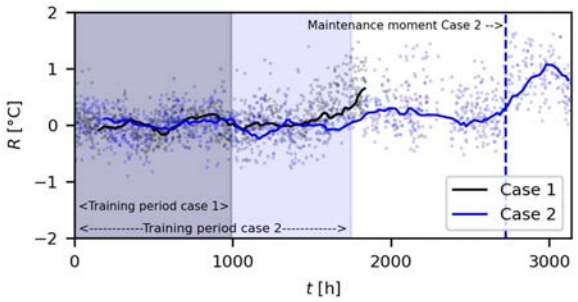


Figure 8. Trajectories of diesel engine temperature residuals with failure at the end ($M = 0.06$, $P = 0.92$, $T = 0.57$, $S = 0.52$)

yielding a score of 0.52.

The monotonicity of the trends is extremely low due to low-frequency oscillations over time. According to Heek (2021), this is caused by unobserved effects in the engines, probably caused by "*minor maintenance actions*", which again can be considered as confounding factors. P is relatively high due to the small difference between end values (i.e. 0.5 and 0.6 respectively), and a weak trendability is observed due to the increase in data points near the end.

However, the actual meaning of these high metrics is disputable. Heek (2021) described that in case 2, maintenance was performed around $t = 2700$h, as indicated by the vertical bar in Fig. 8. After this moment, the residuals make a jump of 1°C. This behavior is also visible when evaluating the third case study in which no failure was observed, shown in Fig. 9. Similar to case 2, maintenance was performed after which residuals make a jump (of approximately 0.6°C) as indicated by the vertical bar around $t = 1200$h.

Again, this yields a gradually increasing trend, but it is unrelated to degradation. Heek (2021) proposed to retrain the regression model after maintenance is performed to reduce the number of false positives, which may provide a solution for

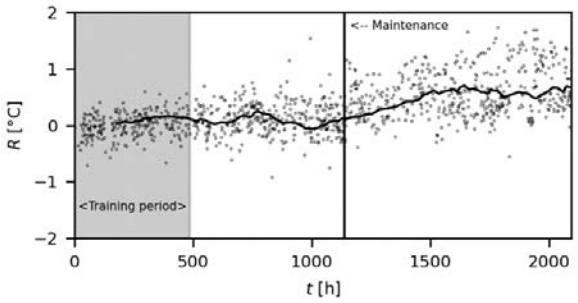


Figure 9. Trajectories of diesel engine temperature residuals without failures

anomaly detection. However, this way the physical relations and meaning of the residuals is inconsistent, i.e. there is no clear link between the residuals and the quantitative amount of damage. This makes it difficult to isolate maintenance actions (or other external confounding factors) from degradation, i.e. the signal-to-noise of the residuals can be considered to be low.

Real-time monitoring the residuals offers potential for early fault detection. However, similar to the ETF case study, it is still extremely complicated to implement prognostics due to the limited number of run-to-failure trajectories and a low signal-to-noise ratio due to limited physical understanding of the data. An additional problem in this case, is that the lack of physical understanding of the residuals complicate threshold definition, and without a threshold it is impossible to estimate the RUL.

2.4. Discussion and Conclusion Case Studies

A first observation of the case studies is that a good direct measurement of the degradation severity (e.g. crack length) can be considered as a perfect prognostic feature. As degradation is in general an irreversible process, it is monotonic and yields perfect trendability. Also thresholds can be defined based on phyical knowledge, yielding perfect prognosability. However, in practice it is complicated or impossible to obtain direct measurements of the actual degradation severity and real-time sensors should provide monotonic, prognosable and trendable run-to-failure trajectories. It is found to be relatively easy to extract such trajectories from well-defined benchmark data sets with labeled data (i.e. milling data set (Agogino & Goebel, 2007)) or many historical run-to-failures (i.e. C-MAPSS (Saxena & Goebel, 2008)).

However, both real-world case studies suffered from two main issues: 1) the number of failures is extremely low, or even non-existent, and 2) low prognostic metrics due to a low signal-to-noise ratio between the health indicator (i.e. the signal) and other confounding factors such as maintenance actions and environments (the noise). The low number of failures, with low prognostic potential in the derived feature, make training of data-driven prognostic algorithms infeasible. Furthermore, the low signal-to-noise ratio makes it complicated to distinguish nominal operating conditions from faulty behavior, such that estimation of the onset of degradation, as well as identifying the degradation trends are difficult.

Note that the latter issue is mainly contributed to confounding factors affecting system behavior, originating from varying operating conditions. This shows the main difference between the benchmark and real-world data sets: in benchmark data sets, either an experimental setup or a simulation is used to generate data, in which external factors influencing system behavior can easily be excluded (lab experiment) or by definition do not exist (simulation). However, in real-world cases

such factors are not always well understood or measured such that they are not included in the developed models. Note that available prognostic algorithms can work if data requirements are met (e.g. as found by van der Velde et al. (2023), chances are higher for assets in large fleets, and requirements are less strict in constant operating conditions (Tiddens et al., 2023)), but a solution needs to be found for these complex real-world cases in varying conditions.

3. Proposed Framework

The lack of run-to-failures and the lack of physical understanding of the monitored signal complicated prognostics in the real-world case studies. Following the decision framework proposed by Tiddens et al. (2023), two solutions are possible: 1) improving the data set or 2) improving the system knowledge. Collecting more run-to-failures is not a realistic option as the failures are critical and therefore prevented by performing preventive maintenance actions. Worldwide data sharing may help (Coelho et al., 2022), especially for similar assets existing in large fleets, such as industrial machinery (Peng et al., 2022) and wind turbines (Li et al., 2021). However, data sharing is often complicated due to standardization and different data structures (Coelho et al., 2022). Furthermore, the military applications bring additional challenges regarding confidentiality of data, but also companies may not be keen on sharing data as they can consider it as intellectual property.

Therefore, the focus in development of the framework (Fig. 10) is on improving system knowledge. System knowledge can be improved in two ways: 1) learning the relation between usage and the degradation rates (i.e. with PoF-models) and 2) and quantifying the relation between measured signals and damage severities. The first part (part I in Fig. 10) is already explored by previous work of the authors (Tinga, 2013a; Keizers et al., 2021, 2022). The second part (part II in 10) focuses on improved quantitative diagnostics to obtain features with higher prognostic metrics, as found to be required for the real-world case studies discussed in section 2.

Degradation can vary heavily between assets used in varying operational conditions (Tiddens et al., 2023), and in absence of historical run-to-failures, the quantitative relation between usage and degradation is essential for accurate prognostics. Therefore, PoF-models are used in part I of the framework. Such models are often tuned for prognostics of specific assets with Bayesian filters (Jouin et al., 2016). However, in literature the effect of actual loading conditions is often simplified, e.g. by substituting loads with a constant model parameter (Zio & Peloni, 2011). This takes away one of the main strengths of a PoF-model, as handling the loads as a model parameter only makes extrapolation of the latest trend possible, yielding wrong RUL for changing future usage profiles, as was shown in Keizers et al. (2021).

To achieve PoF-based prognostics (i.e. part I of the framework) it is proposed to use the method described in Keizers et al. (2021), taking loads as separate input for a Bayesian filter and for prognostics. Loads are first monitored (for $t \leq t_p$, with t_p the time of prediction) to update the PoF-model. Then, expected future loads (for $t > t_p$) are substituted in the updated PoF-model for prognostics (see the lower input of the PoF-model in Fig. 10). This enables RUL prediction based on expected future operating conditions, or adaptation of system usage to extend the RUL.

A conceptual example can be given in the form of the Apache ETF case: it is observed that the ETF decreases faster in sandy environments Vos (2019), which can be explained by increased wear of vanes, blades and seals due to the increased number of sliding sand particles over the components. Erosive wear is already studied for decades (Sundararajan, 1991) and can be modeled by Archard law (Archard, 1953) or by more detailed empirical models that also take characteristics (e.g. hardness, size) of sand particles into account (Gülich, 2020). Such models can be used to estimate degradation rates in specific (and varying) operational and environmental conditions.

The update of the PoF-model requires corresponding degradation measurements. In the studies described in Keizers et al. (2021) and Keizers et al. (2022) direct condition measurements (i.e. measurement of the parameter calculated by the PoF-model) were assumed, which can be the available in some practical applications. For example, in case of fatigue crack growth, DC Potential Drop Methods can measure crack lengths directly (Bär, 2020) and in case of corrosion, electrochemical measurements can measure corrosion rates directly (Homborg et al., 2014). Such direct condition measurements are preferred, as they can be considered to be a perfect prognostic metric as discussed in section 2. However, in many practical applications, such as the real-world case studies of section 2, these types of direct condition measurements are expensive or impossible to obtain. Therefore, monitoring options are used that measure (indirect) consequences of the actual degradation, e.g. vibrations or temperatures.

Here, part II of the framework is introduced. It considers a quantitative diagnostic block, linking indirect condition measurements in specific operating conditions (the right-side input of the block) to the direct condition (i.e. damage severity). However, as illustrated by the case studies in section 2 the features derived from the real-world data sets have low prognostic metrics, and are unlabeled. Therefore, the relation between measured data and degradation severity is unknown, and quantitative diagnostic algorithms cannot be trained. Experimental set-ups could help to gather training data to learn this relation, but it is economically infeasible to collect data for all fault types and locations in all possible operating conditions (Sawalhi & Randall, 2008). Here, the second way of including system knowledge is relevant, which is positioned below the quantitative diagnostic block in Fig. 10. By introducing faults

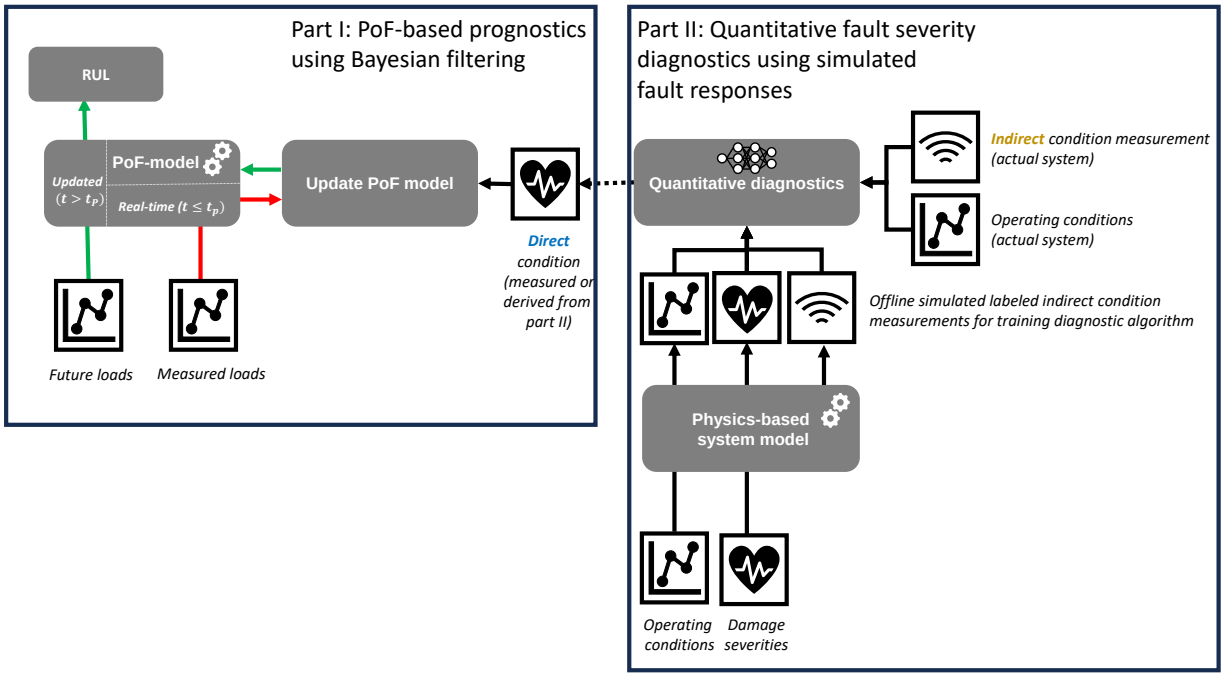


Figure 10. Proposed framework

in physics-based system models and (virtual) measuring the system response of degradation, a better understanding of the effect of faults with varying sizes on measured signals in different operating conditions can be obtained. In the conceptual example of erosive wear in the Apache components, a system model should reveal the efficiency loss for different amounts of material removal. Subsequently, the required additional power, and forthcoming turbine gas temperature, need to be modeled in different operational conditions.

Indeed, developing such a model brings major challenges. First, building an extensive physics-based model for each component or system is time-consuming and expensive. Second, damage induces responses in different physical domains (i.e. mechanical wear yields increased temperature). For these reasons, a modular modeling method that is relatively easy to reuse, adapt and scale, and which is applicable in multiple domains is proposed. Bond graphs are such models (Mkadara et al., 2021). It will be profitable to develop models of standard equipment (e.g. bearings) that can be easily reused in other system models, accelerating development for new machinery.

As an example, Nakhaeinejad & Bryant (2011) showed that different types of bearing faults and their vibration response can be modeled using these types of models. Note that bond graphs are proposed more often in hybrid prognostic frameworks for determining residuals from a nominal system model (e.g. (Medjaher & Zerhouni, 2013)) or for tracking faulty parameters (e.g. Borutzky (2020)), but the link with an actual PoF-model and its corresponding direct condition measurement is missing, limiting prognostic capabilities in cases of varying operating conditions.

To conclude, part II of the framework can help to create a quantitative damage assessment, acting as an input for part I in the likely scenario where direct condition measurements are unavailable. Subsequently, prognostics can be performed, and the RUL can be predicted based on assumed usage profiles. This not only improves prognostic performance in cases of varying operating conditions, but also offers the possibility to adapt usage profiles to extend the RUL.

4. Discussion and Conclusion

The paper showed that prognostic metrics of data from real-world data sets are extremely low compared to benchmark data sets. The main issues observed are unavailability of run-to-failure trajectories and low signal-to-noise ratios of the available trajectories due to always present confounding factors. As a consequence, developed data-driven prognostic methods are often not applicable in practical cases.

The criticality of discussed real-world cases make it unlikely that much run-to-failure data will be collected in the future, so the proposed solution is defined in a framework based on enhancing and utilizing system knowledge. The limited un-

derstanding of measured signals in the real-world cases make trending complicated, so system models with induced damage are proposed to increase understanding of the effects of damage on measured signals. A quantitative diagnostic algorithm can improve the signal-to-noise ratio of measured signals, providing the input for a Bayesian filter that quantifies the relation between usage and degradation rates. Subsequently, prognostics can be performed based on expected system usage.

The framework has strict requirements on system knowledge (i.e. PoF, load monitoring, system models with induced damage), but low data requirements as no historical run-to-failures are needed. To accelerate development for application to new systems, usage of modular system models such as bond graphs is proposed. The framework still needs to be implemented and validated for a real application and the strict requirements on system knowledge requires investments. However, it enables better RUL predictions (or RUL extension by usage adaptation) which can yield great benefits. Before moving to complex cases such as the Apache ETF or the marine diesel engine, it is proposed to start with relatively simple standard equipment such as bearings to validate the benefits of the method. This will be presented in a future publication.

ACKNOWLEDGEMENTS

This work is carried out as a part of the PrimaVera Project, funded by the NWO under grant agreement NWA.1160.18.238. The authors thank the NLR and Royal Netherlands Navy gratefully for making their case studies available for evaluation.

REFERENCES

Agogino, A., & Goebel, K. (2007). *Milling data set.* Retrieved from `https://data.nasa.gov/Raw-Data/Milling-Wear/vjv9-9f3x/data`

Alomari, Y., Andó, M., & Baptista, M. (2023). Advancing aircraft engine rul predictions: an interpretable integrated approach of feature engineering and aggregated feature importance. *Scientific Reports*, *13*(1), 13:13466. doi: https://doi.org/10.1038/s41598-023-40315-1

Archard, J. (1953). Contact and rubbing of flat surfaces. *Journal of Applied Physics*, *24*(8), 981 - 988. doi: https://doi.org/10.1063/1.1721448

Baral, T., Saraygord Afshari, S., & Liang, X. (2023). Residual life prediction of aluminum alloy plates under cyclic loading using an integrated prognosis method. *Transactions of the Canadian Society for Mechanical Engineering*, *47*(5), 1-12. doi: https://doi.org/10.1139/tcsme-2023-0010

Borutzky, W. (2020, Jan). A hybrid bond graph model-based - data driven method for failure prognostic. *Procedia Manufacturing*, *42*, 188-196. (International Conference on Industry 4.0 and Smart Manufacturing (ISM 2019)) doi: https://doi.org/10.1016/j.promfg.2020.02.069

Bär, J. (2020). Crack detection and crack length measurement with the dc potential drop method–possibilities, challenges and new developments. *Applied Sciences*, *10*(23), 8559. doi: https://doi.org/10.3390/app10238559

Calabrese, F., Regattieri, A., Bortolini, M., Gamberi, M., & Pilati, F. (2021). Predictive maintenance: A novel framework for a data-driven, semi-supervised, and partially online prognostic health management application in industries. *Applied Sciences*, *11*(8), 3380. doi: https://doi.org/10.3390/app11083380

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5. doi: https://doi.org/10.3390/data6010005

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022, Jan). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961. doi: https://doi.org/10.1016/j.ress.2021.107961

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018, Sept). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, *307*, 72-77. doi: https://doi.org/10.1016/j.neucom.2018.03.067

Coble, J. B. (2010). *Merging data sources to predict remaining useful life – an automated method to identify prognostic parameters* (Unpublished doctoral dissertation). University of Tennessee.

Coelho, L. B., Zhang, D., Ingelgem, Y. V., Steckelmacher, D., Nowé, A., & Terryn, H. (2022, Jan). Reviewing machine learning of corrosion prediction in a data-oriented perspective. *Materials Degradation*, *6*, 8. doi: https://doi.org/10.1038/s41529-022-00218-4

de Pater, I., Reijns, A., & Mitici, M. (2022, May). Alarm-based predictive maintenance scheduling for aircraft engines with imperfect remaining useful life prognostics. *Reliability Engineering & System Safety*, *221*, 108341. doi: https://doi.org/10.1016/j.ress.2022.108341

Eker, C. F., O.F., & Jennions, I. (2012). Major challenges in prognostics: Study on benchmarking prognostics datasets. In *Proceedings of the european conference of the phm society 2012* (Vol. 1, p. 1-8). doi: https://doi.org/10.36001/phme.2012.v1i1.1409

Elattar, H., Elminir, H., & Riad, A. e.-d. (2016, June). Prognostics: a literature review. *Complex Intelligent Systems*, *2*, 125-154.

Fernandes, M., Chorchaco, J. M., & Marreiros, G. (2022, Mar). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. *Applied Intelligence*, *52*, 14246-14280. doi: https://doi.org/10.1007/s10489-022-03344-3

Ferreira, C., & Gonçalves, G. (2022, April). Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems*, *63*, 550-562. doi: https://doi.org/10.1016/j.jmsy.2022.05.010

Guo, J., Li, Z., & Li, M. (2020). A review on prognostics methods for engineering systems. *IEEE Transactions on Reliability*, *69*(3), 1110-1129. doi: https://doi.org/10.1109/TR.2019.2957965

Gálvez, A., Diez-Olivan, A., Seneviratne, D., & Galar, D. (2021). Fault detection and rul estimation for railway hvac systems using a hybrid model-based approach. *Sustainability*, *13*(12), 6828. doi: https://doi.org/10.3390/su13126828

Gülich, J. F. (2020). *Centrifugal pumps*. Springer.

Heek, D. (2021). *A data-driven condition monitoring approach for the main bearings of a marine diesel engine.* https://research.tue.nl/en/studentTheses/a-data-driven-condition-monitoring-approach-for-the-main-bearings. (MSc thesis, Eindhoven University of Technology)

Homborg, A., Leon Morales, C., Tinga, T., de Wit, J., & Mol, J. (2014, Aug). Detection of microbiologically influenced corrosion by electrochemical noise transients. *Electrochimica Acta*, *136*, 223-232. doi: https://doi.org/10.1016/j.electacta.2014.05.102

Jouin, M., Gouriveau, R., Hissel, D., Péra, M.-C., & Zerhouni, N. (2016, May). Particle filter-based prognostics: Review, discussion and perspectives. *Mechanical Systems and Signal Processing*, *72-73*, 2-31. doi: https://doi.org/10.1016/j.ymssp.2015.11.008

Keizers, L. S., Loendersloot, R., & Tinga, T. (2021). Unscented kalman filtering for prognostics under varying operational and environmental conditions. *International Journal of Prognostics and Health Management*, *12*(2), 1-20. doi: https://doi.org/10.36001/ijphm.2021.v12i2.2943

Keizers, L. S., Loendersloot, R., & Tinga, T. (2022). Atmospheric corrosion prognostics using a particle filter. In *Book of extended abstracts for the 32nd european safety and reliability conference.* doi: https://doi.org/10.3850/978-981-18-5183-4_r22-08-170-cd

Kumar, S., Kolekar, T., Kotecha, K., Patil, S., & Bongale, A. (2022, Jan). Performance evaluation for tool wear prediction based on bi-directional, encoder–decoder and hybrid long short-term memory models. *International Journal of Quality Reliability Management*, *ahead-of-print*, 1551-1576. doi: https://doi.org/10.1108/IJQRM-08-2021-0291

Li, Y., Jiang, W., Zhang, G., & Shu, L. (2021, June). Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data. *Renewable Energy*, *171*, 103-115. doi: https://doi.org/10.1016/j.renene.2021.01.143

Lu, S., Zhu, Y., Liu, S., & She, J. (2022). A tool wear prediction model based on attention mechanism and indrnn. In *2022 international joint conference on neural networks (ijcnn)* (p. 1-7). doi: https://doi.org/10.1109/IJCNN55064.2022.9889794

Lukens, S., Rousis, D., Baer, T., Lujan, M., & Smith, M. (2022). Data quality scorecard for assessing the suitability of asset condition data for prognostics modeling. In *Proceedings of the annual conference of the phm society* (Vol. 14, p. 1-15). doi: https://doi.org/10.36001/phmconf.2022.v14i1.3188

Medjaher, K., & Zerhouni, N. (2013, Jan). Framework for a hybrid prognostics. In (Vol. 33, p. 91-96). doi: https://doi.org/10.3303/CET1333016

Mkadara, G., Maré, J.-C., & Paulmann, G. (2021). Methodology for model architecting and failure simulation supported by bond-graphs—application to helicopter axial piston pump. *Sustainability*, *13*(4), 1863. doi: https://doi.org/10.3390/su13041863

Mulders, M., & Haarman, M. (2017). *Predictive maintenance 4.0 predict the unpredictable* (Tech. Rep.). https://www.pwc.nl/nl/assets/documents/pwc-predictive-maintenance-4-0.pdf. (Retrieved on 21-02-2024)

Nakhaeinejad, M., & Bryant, M. (2011). Dynamic modeling of rolling element bearings with surface contact defects using bond graphs. *Journal of Tribology*, *133*(1), 011102. doi: https://doi.org/10.1115/1.4003088

Peng, D., Liu, C., & Gryllias, K. (2022). A transfer learning-based rolling bearing fault diagnosis across machines. In *Annual conference of the phm society* (Vol. 14, p. 1-9). doi: https://doi.org/10.36001/phmconf.2022.v14i1.3257

Pinciroli, L., Baraldi, P., & Zio, E. (2023, June). Maintenance optimization in industry 4.0. *Reliability Engineering & System Safety*, *234*, 109204. doi: https://doi.org/10.1016/j.ress.2023.109204

Pugalenthi, K., Park, H., Hussain, S., & Raghavan, N. (2021, sept). Hybrid particle filter trained neural network for prognosis of lithium-ion batteries. *IEEE Access*, *9*, 135132-135143. doi: https://doi.org/10.1109/ACCESS.2021.3116264

Ramasso, E., & Saxena, A. (2014). Performance benchmarking and analysis of prognostic methods for cmapss datasets. *International Journal of Prognostics and Health Management*, *5*(2), 1-15. doi: https://doi.org/10.36001/ijphm.2014.v5i2.2236

Sawalhi, N., & Randall, R. (2008). Simulating gear and bearing interactions in the presence of faults: Part i. the combined gear bearing dynamic model and the simulation of localised bearing faults. *Mechanical Systems and Signal Processing*, *22*(8), 1924-1951. doi: https://doi.org/10.1016/j.ymssp.2007.12.001

Saxena, A., & Goebel, K. (2008). *C-mapss data set.*

Sheng, R., & Zhu, X. (2020, Dec). Tool wear assessment approach based on the neighborhood rough set model and nearest neighbor model. *Shock and Vibration*, *2020*, 1-15. doi: https://doi.org/10.1155/2020/8876187

Sun, J., Zuo, H., Wang, W., & Pecht, M. G. (2014). Prognostics uncertainty reduction by fusing on-line monitoring data based on a state-space-based degradation model. *Mechanical Systems and Signal Processing*, *45*(2), 396-407. doi: https://doi.org/10.1016/j.ymssp.2013.08.022

Sundararajan, G. (1991). A comprehensive model for the solid particle erosion of ductile materials. *Wear*, *149*(1), 111-127. doi: https://doi.org/10.1016/0043-1648(91)90368-5

Tiddens, W., Braaksma, J., & Tinga, T. (2023). Decision framework for predictive maintenance method selection. *Applied Sciences*, *13*(3), 2021. doi: https://doi.org/10.3390/app13032021

Tinga, T. (2013a, July). Predictive maintenance of military systems based on physical failure models. *Chemical engineering transactions*, *33*, 295-300. doi: https://doi.org/10.3303/CET1333050

Tinga, T. (2013b). *Principles of loads and failure mechanisms. applications in maintenance, reliability and design*. Springer.

Tinga, T., Wubben, F., Tiddens, W. W., Wortmann, H., & Gaalman, G. (2021). Dynamic maintenance based on functional usage profiles. , *27*(1), 21-42. doi: https://doi.org/10.1108/JQME-01-2019-0002

van der Velde, R., Moerkerken, A., Hofstraat, K., Rosier, M., Haarman, M., de Klerk, P., ... Nedelcheva, Y. (2023). *Digital trends in maintenance* (Tech. Rep.). `https://www.pwc.nl/en/evenementen/digital-trends-in-maintenance.html`. (Retrieved on 01-03-2024)

Virkler, D., Hillberry, B., & Goel, P. (1979). The statistical nature of fatigue crack propagation. *Journal of Engineering Materials and Technology*, *101*(2), 148-153.

Vos, P. (2019). *Engine condition trend monitoring for apache turboshaft engines* (Tech. Rep.). NLR. (Classified)

Wang, J., Wang, P., & Gao, R. (2015, July). Particle filter for tool wear prediction. *Journal of Manufacturing Systems*, *36*, 35-45. doi: https://doi.org/10.1016/j.jmsy.2015.03.005

Wang, P., & Gao, R. (2016). Stochastic tool wear prediction for sustainable manufacturing. *Procedia CIRP*, *48*, 236-241. doi: https://doi.org/10.1016/j.procir.2016.03.101

Yan, J., Meng, Y., Lu, L., & Guo, C. (2017). Big-data-driven based intelligent prognostics scheme in industry 4.0 environment. In *2017 prognostics and system health management conference (phm-harbin)* (p. 1-5). doi: https://doi.org/10.1109/PHM.2017.8079310

Zhou, Y., & Sun, W. (2020, May). Tool wear condition monitoring in milling process based on current sensors. *IEEE Access*, *8*, 95491-95502. doi: https://doi.org/10.1109/ACCESS.2020.2995586

Zio, E., & Peloni, G. (2011). Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering & System Safety*, *96*(3), 403-409. doi: https://doi.org/10.1016/j.ress.2010.08.009

BIOGRAPHIES

Luc S. Keizers received his MSc degree in Mechanical Engineering at the University of Twente in 2020. He graduated on "Structural Fatigue Analysis using Flexible Multibody Dynamics" in the research chair of Applied Mechanics and Data Analysis. Shortly after, he started his PhD at the same university in the chair of Dynamics Based Maintenance in collaboration with the Netherlands Royal Navy. His research interest is in prognostics that combine his engineering background with data-driven models.

Richard Loendersloot received a MSc degree in Mechanical Engineering (2001) and a did his PhD research at the University of Twente, on thermoset resin flow processes through textile reinforcements during composite production process Resin Transfer Moulding and obtained his PhD degree in 2006. He worked in an engineering office on high-end FE simulations of a variety mechanical problems to return to the University of Twente in 2008 as assistant professor for Applied Mechanics. His research started to focus on vibration based structural health and condition monitoring, being addressed in both research and education. He became part of the research chair Dynamics Based Maintenance upon its initiation in 2012. His research covers a broad range of applications: from rail infra structure monitoring, to water mains condition inspection and aerospace health monitoring applications, using both structural dynamics and ultrasound methods. He is involved in a number of European and National funded research projects. He became associate professor in 2019, currently in charge of the daily lead of the Dynamics Based Maintenance group.

Tiedo Tinga is a full professor in dynamics based maintenance at the University of Twente since 2012 and full professor Life Cycle Management at the Netherlands Defence Academy since 2016. He received his PhD degree in mechanics of materials from Eindhoven University in 2009. He is chairing the smart maintenance knowledge center and leads a number of research projects on developing predictive maintenance concepts, mainly based on physics of failure models, but also following data-driven approaches.

Towards a Probabilistic Fusion Approach for Robust Battery Prognostics

Jokin Alcibar[1], Jose I. Aizpurua[2,3], and Ekhi Zugasti[4]

[1,2,4] *Electronics & Computer Science Department, Mondragon University, Spain*
jalcibar@mondragon.edu
ezugasti@mondragon.edu

[3] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
jiaizpurua@mondragon.edu

ABSTRACT

Batteries are a key enabling technology for the decarbonization of transport and energy sectors. The safe and reliable operation of batteries is crucial for battery-powered systems. In this direction, the development of accurate and robust battery state-of-health prognostics models can unlock the potential of autonomous systems for complex, remote and reliable operations. The combination of Neural Networks, Bayesian modelling concepts and ensemble learning strategies, form a valuable prognostics framework to combine uncertainty in a robust and accurate manner. Accordingly, this paper introduces a Bayesian ensemble learning approach to predict the capacity depletion of lithium-ion batteries. The approach accurately predicts the capacity fade and quantifies the uncertainty associated with battery design and degradation processes. The proposed Bayesian ensemble methodology employs a stacking technique, integrating multiple Bayesian neural networks (BNNs) as base learners, which have been trained on data diversity. The proposed method has been validated using a battery aging dataset collected by the NASA Ames Prognostics Center of Excellence. Obtained results demonstrate the improved accuracy and robustness of the proposed probabilistic fusion approach with respect to (i) a single BNN model and (ii) a classical stacking strategy based on different BNNs.

1. INTRODUCTION

Batteries are key components in the transition towards a sustainable carbon-free economy. In this transition, the development of remaining useful life (RUL) prediction of batteries is a crucial activity. The accuracy and reliability of the RUL prediction models is essential to build trust in the predictions (Liu et al., 2023). In this context, robust and reliable battery prognostics models support the development of accurate monitoring strategies and cost-effective solutions.

The estimation of the state-of-health (SOH) of batteries is a key activity for the design of RUL prognostics models. SOH-based prognostics models focus on capturing the run-to-failure ageing dynamics and battery health state estimation (Toughzaoui et al., 2022). It is frequently used to determine age-related degradation that reduces energy capacity and rises safety risks, including overheating and explosions (Wang et al., 2022). Therefore, accurate SOH monitoring and forecasting are key activities to design and operate safe, reliable and effective battery-powered systems (H. Zhao et al., 2023).

SOH estimation is an ongoing area of research (Yang, Chen, Chen, & Huang, 2023). SOH refers to the ratio of the current maximum capacity relative to its original specified capacity (X. Zhao, Wang, Li, & Miao, 2024). SOH can be quantified through different factors, including resistance and maximum power. However, discharge capacity is the most common definition (Vanem, Salucci, Bakdi, & Alnes, 2021), and this is adopted in this research.

Recent data-driven approaches have focused on modeling the capacity degradation of lithium-ion (Li-ion) batteries. (Lee, Kwon, & Lee, 2023) used convolutional neural network (CNN) to estimate the future SOH value of Li-ion batteries, transforming the capacity degradation data into two-dimensional images. Estimates of the SOH and RUL are commonly found together in the literature. For example, (Toughzaoui et al., 2022) developed a CNN-LSTM architecture, and (Wei & Wu, 2023) presented a graph CNN complemented by dual attention mechanisms for the estimation of SOH and RUL of batteries. However, due to the variability inherent in battery manufacturing process, it is essential to quantify this uncertainty to ensure robust and reliable prognostics predictions

(Abdar et al., 2021; Nemani et al., 2023).

There are different sources of uncertainty present in the design, operation and maintenance of batteries (Hadigol, Maute, & Doostan, 2015). (Y. Zhang, Zhang, Liu, Feng, & Xu, 2024) introduced a SOH assessment method that estimates uncertainty through the quantile distribution of deep features, which are inferred from a Residual Neural Network (ResNet) architecture. This approach generates SOH values accompanied by confidence intervals. However, the proposed ResNet architecture lacks probabilistic layers, overlooking the uncertainty inherent in the model parameters. (Che et al., 2024) developed a prognostic framework to assess battery aging, using a CNN-LSTM Bayesian neural network. However, this approach limits the uncertainty to the final dense layers, which are the only components modeled probabilistically.

With the aim of capturing uncertainty associated with complex processes, recent studies in the broader machine learning (ML) community have focused on ensembles of probabilistic models. (Fan, Olson, & Evans, 2017) introduced a Bayesian posterior predictive framework for weighting ensemble climate models. (Cobb et al., 2019) present a new ML retrieval method based on an ensemble of Bayesian Neural Networks (BNNs). In this scenario, the overall output from the ensemble is treated as a Gaussian mixture model. However, models are equally weighted with no adaptation to the observed data. (S. Zhang, Liu, & Su, 2022) present a Bayesian Mixture Neural Network (BMNN) for Li-ion battery RUL prediction. The BMNN framework incorporates a Bayesian Convolutional Neural Network as feature extractor and a Bayesian Long Short-Term Memory to learn degradation patterns over time. However, the absence of a weighted model combination limits the analysis of individual model contributions.

Alternatively, (Bai & Chandra, 2023) described a Bayesian ensemble learning framework that uses gradient boosting by combining multiple Neural Networks trained by Markov Chain Monte Carlo (MCMC) sampling. Finally, (Dai, Pollock, & Roberts, 2023) demonstrate the robustness of Bayesian fusion by embedding the Monte Carlo fusion framework within a sequential Monte Carlo algorithm.

In this context, inspired by the use of probabilistic ensemble models to capture model uncertainty, the main contribution of this research is the development of a novel probabilistic model fusion approach for battery SOH predictions. Bayesian convolutional neural networks (BCNNs) are used as base models for SOH prediction, and the fusion approach integrates individual BCNN probabilistic predictions. The fusion strategy balances between precision and reliability of individual predictions, adopting an optimal tradeoff between accuracy and uncertainty of predictions through the proposed stacking approach.

The proposed approach has been compared with (i) individual BCNN models and (ii) fusion strategies focused on stacking of BCNN models using point prediction information. Obtained results confirm that the proposed framework infers accurate, well-calibrated, and reliable probabilistic predictions, which improve predictive performance and contribute to estimate uncertainty in a robust and reliable manner in complex data-driven tasks. The proposed approach has been tested and validated with the publicly available NASA's battery dataset (Saha & Goebel, 2007).

The remainder of this article is organized as follows. Section 2 outlines our probabilistic fusion approach for robust battery prognostics. Section 3 describes a case study to demonstrate the application of our methodology. Section 4 presents and analyzes the results obtained from the case study. Section 5 discusses the implications of these findings. The article concludes with Section 6, summarizing our main conclusions and suggesting avenues for future work.

2. Probabilistic Fusion Approach for Robust Battery Prognostics

The proposed probabilistic fusion framework integrates BCNNs with probabilistic ensemble strategies. The main objective of the integration is to generate accurate predictions with robust uncertainty quantification, thanks to the uncertainty quantification of Bayesian modelling (Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015) and the robustness and accuracy of ensemble strategies (S. Zhang et al., 2022).

The approach is divided into offline and online stages. Starting from a set of battery datasets, in the offline process, data pre-processing and model training steps are completed. In the online process, trained models are stacked in an ensemble model according to computed weight and stacking criteria. The outcome of the approach is a one-step-ahead probabilistic capacity estimate. Figure 1 shows the high-level block diagram of the proposed approach.

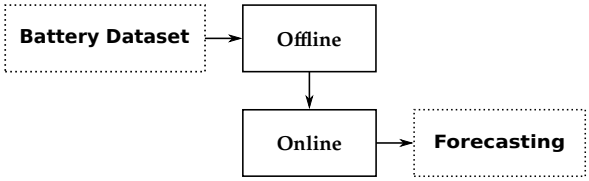


Figure 1. High-level block diagram of the proposed approach.

The high-level concepts in Figure 1 are implemented through the detailed model architecture shown in Figure 2.

The base models are BCNN models, which are trained (offline) through a leave-one-out cross validation (LOOCV) process. The probabilistic results of individual BCNN models are aggregated through a stacking process that includes accuracy and uncertainty metrics. In the testing (online) phase, each BCNN model weights are computed using learned mod-

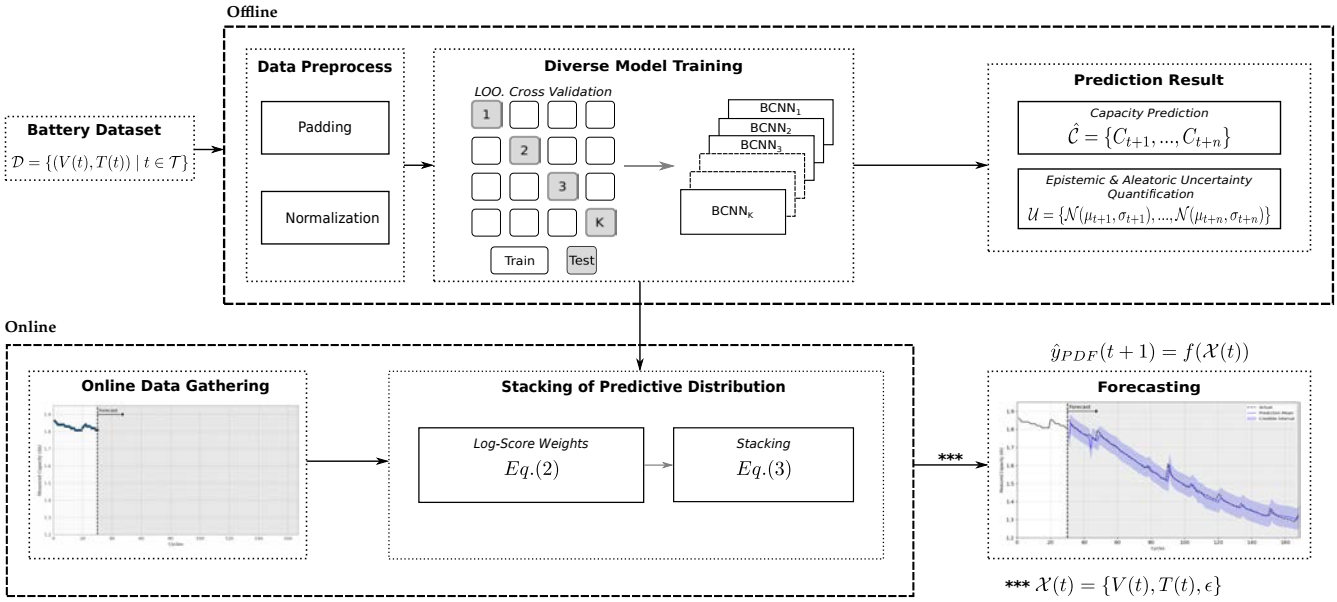


Figure 2. Block diagram of the proposed approach.

els (log-score weights) and the stacking model is designed to combine them and generate a distribution from a mixture model. The following subsections explain in detail the main parts of the approach.

2.1. Offline Phase

During the offline phase, starting from a battery dataset with different run-to-failure trajectories on the same type of batteries, different base models are designed through a training strategy which seeks diversity in the training set to develop complementary predictive models.

2.1.1. Ensemble Base Models: BCNNs

BCNN models are a Bayesian extension of the classical CNN models to include uncertainty associated with parameter estimation. This requires modification of the classical backpropagation algorithm through Bayesian techniques that involves incorporating uncertainty into the model by treating weights as random variables, and applying variational inference to approximate posterior distributions. This results in a more robust model that predicts the complete probability density function (PDF).

Consequently, BCNN models have been selected to improve the robustness and accuracy of model prediction. To this end, BCNNs make use of probabilistic distributions to model parameters and the uncertainty related to their training process, and prior distributions to incorporate previous knowledge, generate uncertainty estimations and mitigate over-fitting (Blundell et al., 2015). In contrast, the classical learning models, e.g. non-Bayesian CNN models, focus on maximum likelihood estimation (MLE) and they overlook prior and posterior distributions. This leads to increasing error and decreasing model robustness in high uncertainty contexts, e.g. out-of-distribution data or manufacturing drifts.

The proposed approach utilizes data pre-processing techniques to standardize the length of discharge cycles through padding. This technique involves repeating the last discharge value until the desired cycle length is reached, ensuring consistent input dimensions for all models. Additionally, normalization is carried out scaling the discharge values between 0 and 1.

The architecture of the BCNN models is shown in Figure 3 defined as follows:

- Input data: the input data for the BCNN is structured in a tensor format. The rows represent data samples of discharge cycles, and columns that correspond to features, such as the voltage and temperature over time. Notably, the input does not include the current discharge as it remains constant in this scenario.
- Convolutional 1D Reparametrization: this layer creates a convolution kernel that is applied to the input data. During the forward pass, kernel and bias parameters are drawn from a Gaussian distribution. It uses the reparameterization estimator to approximate distributions through Monte Carlo trials, integrating over the kernel and bias.
- Global Average Pooling 1D: this layer performs average pooling specifically for temporal data. It reduces the spatial dimensions of the input data to a single value per channel by calculating the average over the temporal dimension.
- Flatten: this layer reshapes input data into a one dimensional array, enabling compatibility between Bayesian

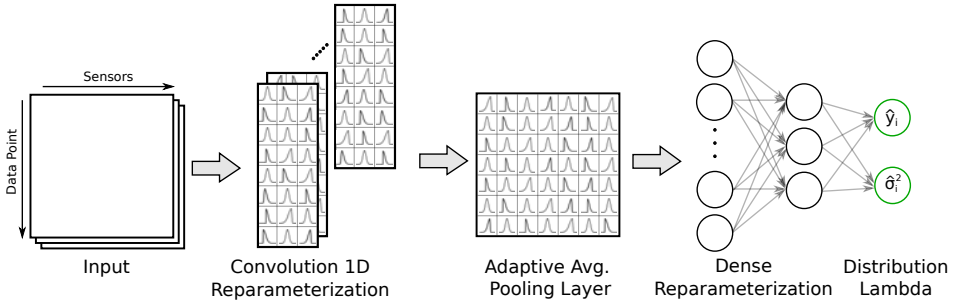


Figure 3. Schematic of the Bayesian convolution neural network.

convolutional layers and Bayesian dense layers.

- Dense Reparameterization: this layer implements a reparameterization estimator for Bayesian variational inference. It implements a stochastic forward pass via sampling from the kernel and bias distributions. This approach improves the robustness of the model, allowing uncertainty estimation in parameter values and supporting probabilistic modeling in deep learning.
- Distribution Lambda: this layer is responsible for producing the final results given the inputs and the learned weights from the previous layers. The output layer consists of two neurons representing the mean, $\hat{y}$ and variance, $\hat{\sigma}^2$, in order to quantify the expected value and its associated uncertainty. To ensure a positive variance, the neuron is activated using an exponential function.

BCNN combines feature extraction capabilities of classical CNN models with the uncertainty quantification of Bayesian theory. The proposed architecture is built using the Bayesian layers of `TensorFlow Probability` in Python (Dillon et al., 2017).

2.1.2. Training for Diversity

Model diversity is a key concept for effective ensemble models (Nam, Yoon, Lee, & Lee, 2021). Accordingly, in this case, the training set for each battery model is modified to learn different battery aging properties. Historical capacity fading data are used to build aging models for each battery in the dataset.

Namely, using the LOOCV strategy, if K run-to-failure trajectories are available, K diverse BCNN models are built changing the training set in each iteration (cf. Figure 2). That is, the model is trained on all batteries except one, which is held as a test set. This process is repeated so that each battery serves as a test set exactly once. Thus, all available data are used for training, maximizing the diversity of training scenarios.

Training the BCNN models through LOOCV strategy, enhances the ability of individual models to generalize across different battery types and manufacturing conditions.

This stage completes the offline training process, which results in a set of BCNN models:

$$\mathcal{M} = \{BCNN_1, BCNN_2, \ldots, BCNN_K\}, \qquad (1)$$

which are used in the subsequent online inference process to build ensemble models.

2.2. Online: Stacking of Predictive Distribution

During the online phase, the proposed stacking of predictive distribution strategy is designed and tested. The proposed approach takes as input individual base models [cf. Eq. (1)] and monitored data up to the prediction instant t, which is used to forecast the probability density function (PDF) of the capacity at $t+1$, $\hat{y}_{PDF}(t+1)$. The objective of the stacking process is to integrate the predictive distributions of different base models and propagate all the information end-to-end.

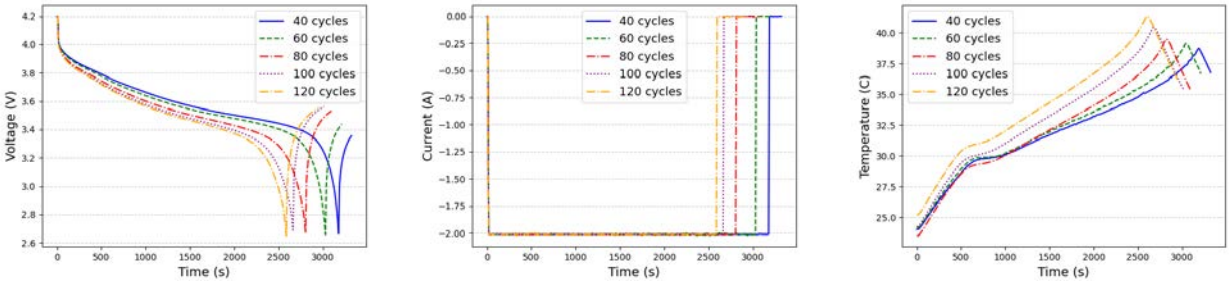
For comparison and benchmarking purposes, an alternative stacking approach is also implemented named stacking of point prediction (cf. Subsection 3.3).

Log-Score Weights

The optimal way to combine a set of Bayesian posterior predictive distributions is by using the logarithmic score (Yao, Vehtari, Simpson, & Gelman, 2018). This method maximizes the average log-likelihood of the observed data, which is a proper scoring rule used to evaluate the accuracy of probabilistic forecasts. It measures the accuracy of a forecast and penalizes overconfidence and underconfidence in the predicted probability. The logarithmic score is defined as follows:

$$\hat{w} = \arg\max_{w} \frac{1}{N} \sum_{i=1}^{N} log \sum_{k=1}^{K} w_k p(y_i \mid y_{-i}, M_k) + \lambda_{reg} \sum_{k=1}^{K} w_k^2 \quad (2)$$

where N denotes the total number of data points and K denotes the total number of base models. The leave-one-out predictive distribution for each model, *i.e.* $p(y_i \mid y_{-i}, M_k)$, is used to compute the model's prediction for the data point i. To avoid overfitting, a regularization term λ_{reg} is added to the likelihood function, penalizing large weights.



(a) Voltage variation (b) Current variation (c) Temperature variation

Figure 4. Feature variations due to an increasing number of discharge cycles in battery #5.

Stacking

Stacking is a method to average point estimates from several models (LeBlanc & Tibshirani, 1996). In its simplest form, it can be seen as a weighted average method. Through the weighted average, it facilitates the construction of ensembles that incorporate predictions from multiple models. In the proposed framework, the goal of weighted average ensemble is to leverage the predictive capabilities of K pre-trained BCNN models [cf. Eq. (1)]. It seeks to mitigate forecasting errors by assigning weights to the linear combination of these models, thereby enhancing the accuracy of predictions.

In the Bayesian framework, stacking extends beyond the limitations of averaging point predictions by combining multiple Bayesian posterior predictive distributions. This approach develops a *stacking model* that leverages the strengths of various predictive models, enhancing overall predictive accuracy. The stacking of the predictive distribution enables the fusion of uncertainties from various models into a unified predictive framework. This approach improves the accuracy of forecasts and offers a comprehensive evaluation of the uncertainty associated with these forecasts, providing advantages across diverse decision-making scenarios. The fundamental equation governing this process is defined as follows:

$$\hat{p}(\tilde{y}|y) = \sum_{k=1}^{K} \hat{w}_k p(\tilde{y}|y, M_k) \tag{3}$$

where $\hat{p}(\tilde{y}|y)$ represents the aggregate probability estimation based on the ensemble model, ω_k denotes the weight assigned to the k-th component within the ensemble, and $p(\tilde{y}|y, M_k)$ refers to the probabilistic forecast generated by each base model, denoted as BCNN_k, given the observed data y.

This probabilistic prediction indicates the likelihood of observing the predicted outcome $\tilde{y}$, dependent on the specific base model employed.

2.2.1. Forecasting

Online forecasting is computed for one-step-ahead predictions. In order to forecast battery capacity at instant $t+1$, previous data until the instant t is used, plus an uncertainty factor expressed as noise:

$$\mathcal{X}(t) = \{V(t), T(t), \epsilon\} \tag{4}$$

where $\{V(t), T(t)\}$ denote the values of voltage and temperature at instant t, and ϵ denotes the Gaussian noise term, $N(0, \sigma)$ with $\sigma = 0.1$, that introduces variability in the progression of X over time.

The one-step-ahead capacity distribution prediction is thus defined as follows:

$$\hat{y}_{PDF}(t+1) = f(\mathcal{X}(t)) \tag{5}$$

where $f(.)$, denotes the designed ensemble model, $\hat{y}_{PDF}(t+1)$ is the distribution of the capacity estimate at $t+1$.

It is possible to perform SOH predictions for longer prediction horizons through a recursive forecasting strategy. However, due to the accumulation of individual forecasting errors, this approach may lead to decrease long-term forecasting performance. Long-term SOH forecasting activities are left open for future work.

This approach allows the model to learn continuously and adapt to changing conditions. Online forecasting is particularly beneficial in environments that require immediate decision making based on the latest available data.

3. CASE STUDY

3.1. Dataset description

The effectiveness of the proposed method has been tested using a battery dataset from the NASA Ames Prognostics Center of Excellence (Saha & Goebel, 2007).

A subset of available battery data has been selected, focusing on batteries #5, #6, #7 and #18. Each battery is operated under various conditions including charging, discharging, and impedance analysis. Throughout the charge and discharge cycles, temperature, current, and voltage were meticulously recorded. During charging, a constant current mode at 1.5 A was maintained until the voltage reached 4.2 V, followed by a switch to constant voltage mode until the current dropped to 20 mA. Discharge cycles involved a constant load mode at 2 A until the voltage levels reached 2.7 V, 2.5 V, 2.2 V and 2.5 V for batteries #5, #6, #7 and #18, respectively. The experiment ended once the battery capacity decreased by 30%. These batteries had a maximum capacity of 2Ah with an end-of-life capacity set at 1.4Ah.

Figures 4(a), 4(b) and 4(c) show the evolution of voltage, current (constant), and temperature measurements with the increment of discharge cycles for the battery #5. Figure 5 shows variations in capacity degradation rates for identical batteries. This is an indicator of uncertainty inherent in the manufacturing process, which affects SOH estimates.

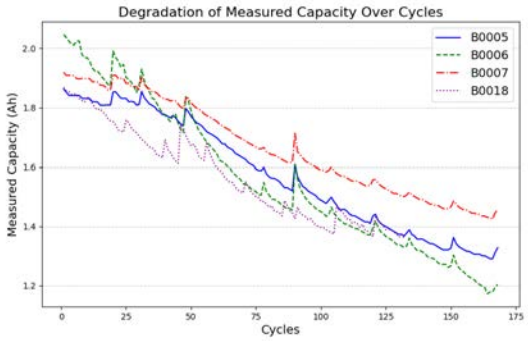


Figure 5. Capacity degradation data of Li-ion batteries.

3.2. BCNN structure and hyperparameters

The design of the base BCNN model structure is developed through experimentation. The BCNN architecture for SOH forecasting is detailed in Table 1, where 'None' is indicative of the batch size. The input for the model comprises 371 data points per discharge cycle, with each point aggregating 3 features: voltage, temperature, and time.

The proposed structure encompasses a total of 1300 trainable parameters, designed to extract features from battery discharge cycle data for forecasting purposes. Figure 3 details the convolutional layer hyperparameters, which includes 16 kernels, each with a dimension of 3, adopting a Laplace distribution for the prior and employing a ReLU activation function. In addition, the model incorporates Bayesian dense layers with 16 units, Adam optimizer, a learning rate of 0.01, and Evidence Lower Bound (ELBO) as its loss function (S. Zhang et al., 2022).

Table 1. BCNN model architecture

Layer	Description	Output Shape	# Param.
-	Input	(None, 371, 4)	0
1	Conv.1D Reparameter.	(None, 369, 16)	416
2	Conv.1D Reparameter.	(None, 368, 8)	528
3	Global Average Pooling	(None, 8)	0
4	Flatten	(None, 8)	0
5	Dense Reparameter.	(None, 16)	288
6	Dense Reparameter.	(None, 2)	68
7	Distribution Lambda	(None,1),(None,1)	0
	Total params: 1300 (5.08 KB)		

3.3. Benchmarking

In order to compare the designed stacking approach with alternative stacking strategies, another stacking approach has been designed using point prediction information instead of the full distribution.

Stacking of Point Prediction

An effective method for determining the weight of each model in the stacking process is by minimizing the leave-one-out mean squared error with a L_2 regularization term, λ_{reg}. The purpose of this term is to penalize large weights, thus preventing overfitting and balancing individual model contributions. The weights are obtained through the following optimization problem:

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} w_k \hat{f}_K^{(-i)}(x_i) \right)^2 + \lambda_{reg} \sum_{k=1}^{K} w_k^2 \quad (6)$$

where $\hat{f}_K^{(-i)}(x_i)$ represents the predicted value of the k-th model, when the i-th observation is left out of the training set. The regularization parameter, λ_{reg}, controls the strength of the regularization applied. To ensure a feasible solution, the weights are restricted to $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$.

Accordingly, the stacking of point prediction approach is defined as follows:

$$\hat{y} = \sum_{k=1}^{K} \hat{w}_k f_k(x|\theta_k) \quad (7)$$

where $\hat{y}$ represents the prediction of the ensemble for the test battery capacity, $\hat{w}_k$ denotes the weight assigned to the k-th battery base model, and $f_k(x|\theta_k)$ is the prediction made by the corresponding base model (BCNN_k).

3.4. Evaluation criteria

The accuracy of the regression is measured by Mean Squared Error, while Negative Log Likelihood assesses model perfor-

mance by quantifying prediction probabilities. Finally, The correctness of probability predictions is assessed through the CRPS.

Mean Square Error (MSE) is a metric for measuring the quality of an estimator. It is a measure of the average squared differences between the estimated values and what is estimated. MSE is calculated by taking the average of the square of the differences between the predicted values and the actual values (Hodson, 2022).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{8}$$

where, n represents the number of observations, Y_i denotes the actual value for the ith observation, and $\hat{Y}_i$ signifies the predicted value for the ith observation.

Coefficient of Determination (R^2) is a metric used to assess the goodness of fit of the model. It provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model (Barrett, 1974).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{9}$$

where, n is the number of observations, Y_i is the actual value, $\hat{Y}_i$ the predicted value for the i-th observation and $\bar{Y}$ the mean of Y. R^2 of 1 implies perfect model predictions, while 0 means no explained variability.

Continuous Ranked Probability Score (CRPS) can be formally expressed as a quadratic measure of discrepancy between the predicted Cumulative Distribution Function (CDF), $F(\cdot)$, and the observed empirical CDF for a given scalar observation y (Zamo & Naveau, 2018):

$$CRPS(F, y) = \int (F(x) - \mathbb{I}(x \geq y_i))^2 dx, \tag{10}$$

where $\mathbb{I}(x \geq y_i)$ is the indicator function, which models the empirical CDF.

To obtain a single score value from Eq. (10), a weighted average is calculated for each individual observation of the test set (Gneiting, Raftery, Westveld, & Goldman, 2005):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N} CRPS(F_i, y_i) \tag{11}$$

where N denotes the total number of predictions.

Negative Log Likelihood (NLL) metric assesses probabilistic models by using the likelihood concept, which indicates how likely the observed data is given model parameters (Bosman & Thierens, 2000). Likelihood ($\mathcal{L}$) is the product of each observation's probability density function (PDF), expressed mathematically as

$$\mathcal{L}(\theta \mid X) = \prod_{i=1}^{N} f(x_i|\theta) \tag{12}$$

where θ denotes model parameters and X includes N data points. NLL is preferred for optimization since minimizing NLL is equivalent to maximizing the log-likelihood, facilitating the discovery of model parameters that best explain the observed data, represented by

$$-\log \mathcal{L}(\theta \mid X) = -\sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{13}$$

Calibration refers to the statistical consistency between the predictive distributions and the actual observations. It represents a joint property of forecasts and empirical data (Jung, Jo, Choo, & Lee, 2022). Namely, it is stated that the model is calibrated if (Kuleshov, Fenner, & Ermon, 2018):

$$\frac{\sum_{t=1}^{T} \mathbb{I}\{y_t \leq F_t^{-1}(p)\}}{T} \rightarrow p \text{ for all } p \in [0, 1] \tag{14}$$

In this expression, T refers to the total number of data points, while the indicator function $\mathbb{I}\{y_t \leq F_t^{-1}(p)\}$ takes a value of 1 when the condition $y_t \leq F_t^{-1}(p)$ is true, and 0 otherwise. Given this condition, y_t express the observed outcome at time t, and $F_t^{-1}(p)$ is the inverse of the CDF for the forecast, evaluated at probability p. Therefore, the condition represents the threshold below which a random sample from the distribution would occur with a probability p.

Sharpness means that the confidence intervals should be optimized for minimal width around a singular value. That is, the goal is to reduce the variance, denoted as $var(F_n)$, of the random variable characterized by the cumulative distribution function F_n (Kuleshov et al., 2018; Tran et al., 2020):

$$sha = \sqrt{\frac{1}{N}\sum_{n=1}^{N} var(F_n)} \tag{15}$$

Table 2. Comparison of different ensemble strategies for different batteries used as test.

	Baseline Model				Benchmarking Ensemble				Proposed Ensemble			
	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$
B0005	0.0007	0.9732	2.3397	0.0183	**0.0002**	**0.9901**	-1.9523	0.0145	0.0003	0.9886	**-2.1001**	**0.0131**
B0006	0.0013	0.9636	8.0947	0.0213	**0.0009**	**0.9753**	-1.8222	0.0183	0.0009	0.9741	**-1.9358**	**0.0178**
B0007	0.0005	0.9696	-0.0409	0.0149	**0.0003**	**0.9814**	**-1.9755**	**0.0145**	**0.0004**	**0.9763**	**-1.9769**	**0.0145**
B0018	0.0013	0.8943	9.0342	0.0223	**0.0010**	**0.9183**	**-1.9478**	**0.0174**	**0.0010**	**0.9141**	**-1.9312**	**0.0178**

4. RESULTS

To evaluate the proposed approach, firstly, different ensemble strategies are compared to evaluate their strengths and identify the most suitable approach. Subsequently, a sensitivity analysis is developed with respect to the contribution of individual base-models to the overall ensemble.

4.1. Probabilistic Ensemble Strategies

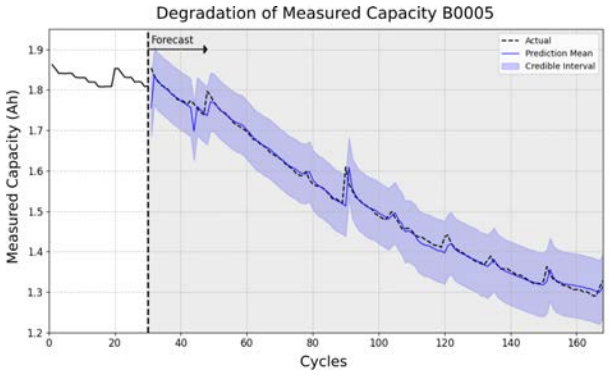
This section focuses on the comparison between (i) the baseline model, *i.e.* BCNN model trained with all available data, (ii) ensemble of point prediction and (iii) proposed ensemble method (cf. Figure 2) to further evaluate the improvement of ensemble strategies over baseline model.

Table 2 presents a comparative analysis in terms of accuracy and probabilistic metrics. This comparison highlights that, for different test scenarios, the ensemble methodologies enhance the performance of the baseline model.

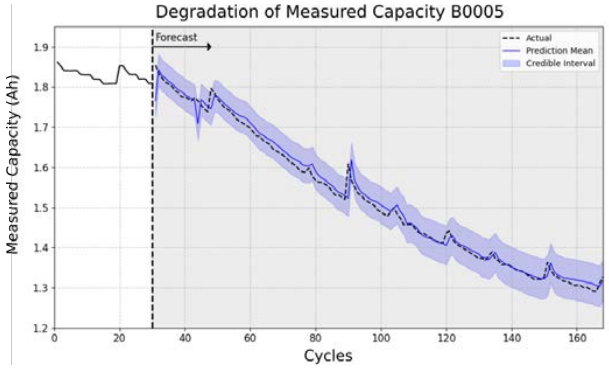
A notable observation from the results in Table 2 is the variance between the proposed ensemble approach (cf. Figure 2) and the benchmarking ensemble model (cf. Subsection 3.3) in specific scenarios. For batteries #5 and #6, the proposed approach exhibited superior outcomes, particularly in probabilistic metrics (NLL and CRPS). This suggests that within a Bayesian framework, prioritizing likelihood maximization, leads to accurately modelling uncertainty, and therefore, it is more advantageous than focusing on MSE minimization (as in Subsection 3.3).

The model optimization criterion has a direct impact on the performance of the tested methods and on the effectiveness of the ensemble approach. However, for batteries #7 and #18, no significant differences were observed between the tested ensemble approaches, which indicates that the results are associated to the prior models. That is, it is possible that the same prior model minimizes the MSE and maximizes the likelihood at the same time.

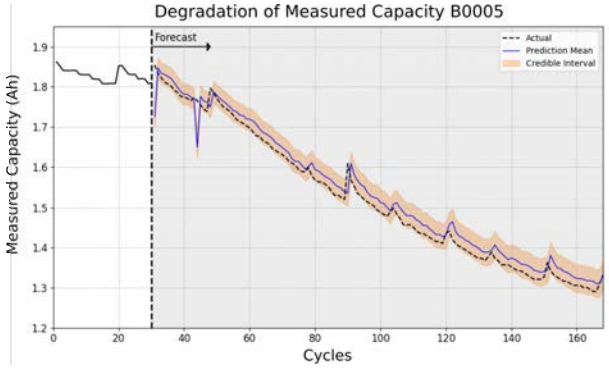
Figure 6(a) shows the comparison between the ensemble model generated by stacking point predictions (cf. Subsection 3.3), Figure 6(b) shows the ensemble model generated through stacking of predictive distributions (cf. Figure 2), and Figure 6(c) shows the individual BCNN trained with the entire dataset, e.g. for the battery #5, train with batteries #6, #7, and #18, and test with #5.



(a) Stacked point prediction method (cf. Subsection 3.3)



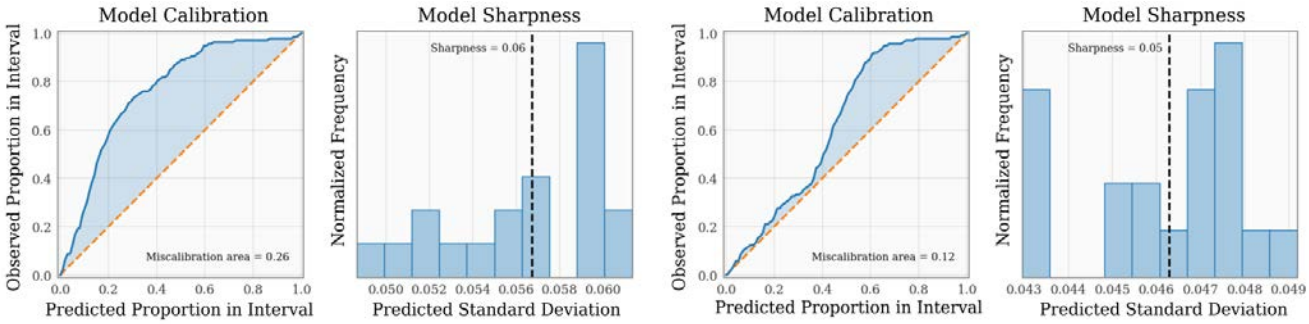
(b) Stacked predictive distribution method (cf. Figure 2)



(c) Baseline model

Figure 6. Battery capacity degradation forecasting results.

It is observed that the ensemble models enhance the performance of baseline model in terms of accuracy and uncertainty



(a) calibration and sharpness for the benchmarking ensemble model

(b) calibration and sharpness for the proposed ensemble model

Figure 7. Evaluation of calibration and sharpness for battery #5.

quantification. This is indicated by the positioning of the ground truth (dashed lines) at the limit of the lower boundary in Figure 6(c), which means that the uncertainty does not accurately cover the observed values. That is, the uncertainty bounds are not well-calibrated, compromising the model's ability to accurately represent the underlying variability in the data and in the model compared to ensemble strategies.

Figure 6(a) shows an improvement in the prediction accuracy. However, it simultaneously introduces a higher level of uncertainty compared to the proposed ensemble method in Figure 6(b). This is reflected in the NLL and CRPS metrics, where the stacking of the predictive distribution demonstrates superior performance (cf. Table 2). Such probabilistic metrics indicate that the model parameters make the observed data more probable, indicating a good fit to the observed data.

The evaluation of the shape of the PDF is a crucial aspect of uncertainty quantification. Accordingly, the calibration and the sharpness assessment of PDFs is performed through a python toolbox for predictive uncertainty quantification (Chung, Char, Guo, Schneider, & Neiswanger, 2021). Figure 7 shows the calibration and sharpness of the analysed ensemble methods designed for probabilistic forecasting for the battery #5.

The calibration plot for the point-prediction ensemble model [cf. Figure 7(a)] reveals a miscalibration area of 0.26, indicating a gap between predicted probabilities and actual outcomes, generally overestimating event probabilities. On the contrary, the proposed ensemble model [cf. Figure 7(b)] shows better calibration with a miscalibration area of 0.12, aligning closer to the ideal, especially in midrange probabilities.

In terms of sharpness, the predictions of the point-prediction based ensemble model have a mean sharpness value of 0.06 and are right-skewed, reflecting higher uncertainty. However, the proposed ensemble model has a mean sharpness value of 0.05, with a slightly left-skewed distribution, indicating more predictions with lower uncertainty and greater confidence.

4.2. Sensitivity of the Ensemble Strategy with Base-Models

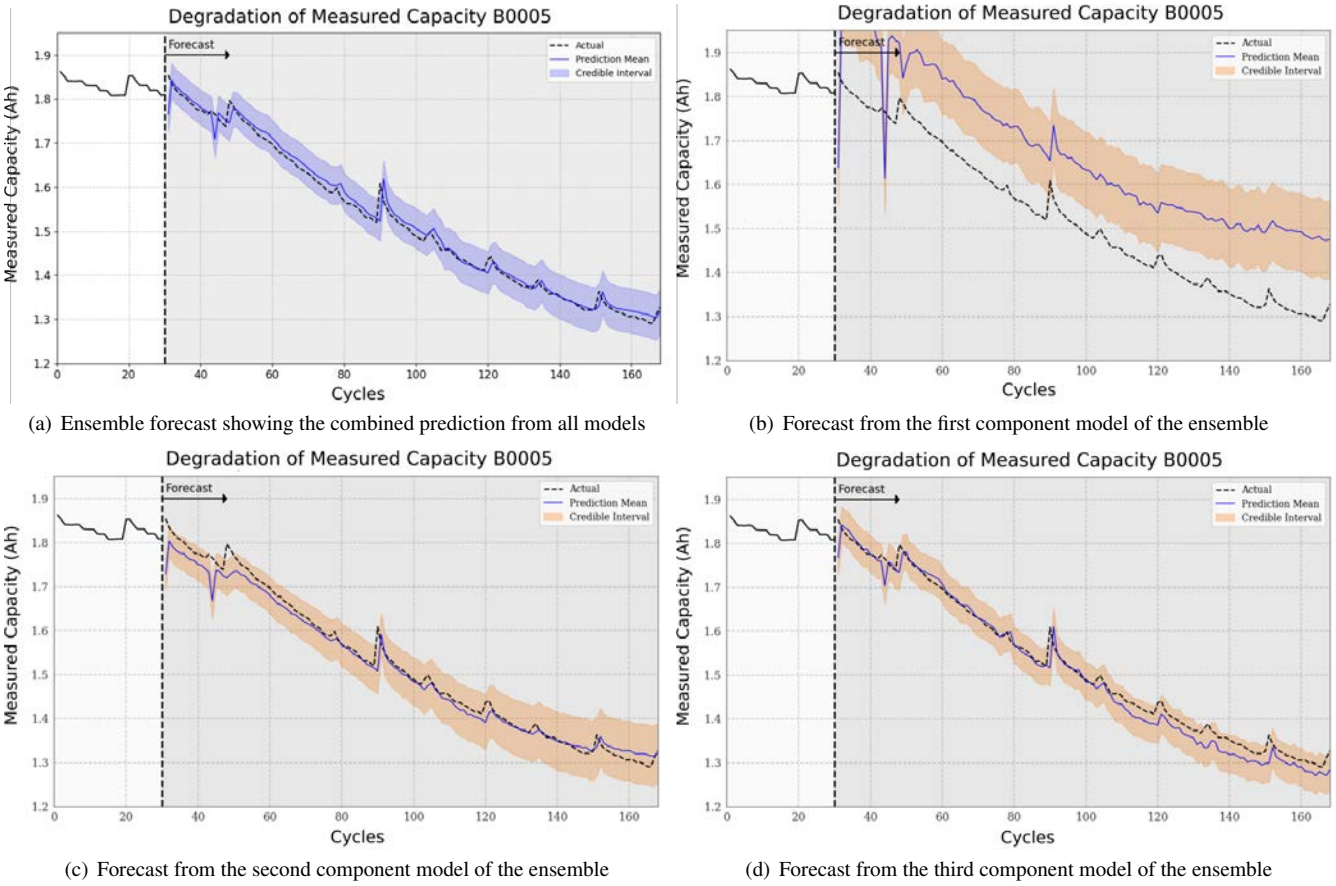
To evaluate the contribution of each individual BCNN model to the ensemble approach, a sensitivity assessment has been performed. Namely, the performance of the different leave-one-out iterations has been evaluated, sequentially training with different battery datasets and testing with the leave-out battery dataset. This has been compared with the proposed ensemble approach results to identify individual contributions from different models. Table 3 displays the obtained results.

Table 3. Performance evaluation of BCNN models and the ensemble approach.

Test[1]	Model	*MSE* (↓)	R^2 (↑)	*NLL* (↓)	*CRPS* (↓)
#5	BCNN [#6,#7][2]	0.0005	0.9802	-1.0707	0.0135
	BCNN [#6,#18]	0.0244	0.1016	19.4417	0.1411
	BCNN [#7,#18]	0.0006	0.9795	-2.0774	0.0132
	Ensemble	**0.0003**	**0.9886**	**-2.1001**	**0.0131**
#6	BCNN [#5,#7]	0.0011	0.9695	3.7012	0.0197
	BCNN [#5,#18]	0.0147	0.5861	0.5852	0.0849
	BCNN [#7,#18]	0.0018	0.9491	-0.7498	0.0252
	Ensemble	**0.0009**	**0.9741**	**-1.9358**	**0.0178**
#7	BCNN [#5,#6]	0.0008	0.9543	-1.5462	0.0166
	BCNN [#5,#18]	0.004	0.7704	2.1996	0.0326
	BCNN [#6,#18]	0.0026	0.854	-1.5735	0.0286
	Ensemble	**0.0004**	**0.9763**	**-1.9769**	**0.0145**
#18	BCNN [#5,#6]	0.0091	0.2534	14.708	0.0833
	BCNN [#5,#7]	0.0041	0.6663	1.5441	0.0459
	BCNN [#6,#7]	0.0013	0.8929	1.8299	0.0213
	Ensemble	**0.0010**	**0.9141**	**-1.9312**	**0.0178**

[1] Battery identifier used for testing.
[2] BCNN [#A,#B]: BCNN trained with batteries #A and #B.

The ensemble BCNN model demonstrates significantly higher accuracy and predictive power than individual BCNN models, as evidenced by its superior performance across multiple metrics. It achieves the lowest MSE in every testing battery, indicating more precise predictions, and the highest R^2 score, showing its ability to explain a greater proportion of variance.



(a) Ensemble forecast showing the combined prediction from all models

(b) Forecast from the first component model of the ensemble

(c) Forecast from the second component model of the ensemble

(d) Forecast from the third component model of the ensemble

Figure 8. Capacity fade forecasting for battery #5 employing an ensemble of BCNN models.

The ensemble model also shows a notable improvement in the NLL metric, suggesting a more reliable uncertainty estimation. Additionally, by achieving the lowest CRPS, it emphasizes its proficiency in probabilistic forecasting and precise uncertainty quantification. Overall, the ensemble method outperforms individual models, highlighting its effectiveness in contexts that require high accuracy and reliability.

Figure 8 presents the forecasts generated by individual models for battery #5 (cf. Table 3). Figures 8(b)-8(d), show individual models and Figure 8(a) shows the combined forecast of the ensemble model.

It can be seen that the ensemble effectively combines the characteristics of models 2 and 3, thereby improving the overall performance of the final forecast of the ensemble.

5. Discussion

The proposed research work demonstrates that the stacking of predictive distributions based on a Bayesian framework improves the accuracy and robustness of predictions compared with stacking of point predictions. Furthermore, it has been observed that the use of an ensemble of BCNN models improves the modeling of uncertainty when compared to relying on a single BCNN model (baseline). However, before drawing definitive conclusions about the application of the proposed solution in real-world applications, further work is necessary testing the robustness, scalability, and sensitivity with respect to noise.

Robustness

Credible intervals reflect the uncertainty associated with the data and the model (cf. Figure 6). The robustness of the proposed approach is therefore directly dependent on model and data uncertainty. The reduction of credible intervals align with the objective of increasing robustness. To this end, increasing the number of observations would reduce the uncertainty attributed to the model, which results in more precise credible intervals. Additionally, employing priors like maximum entropy priors or weakly informative priors may further tighten credible intervals, thereby improving the reliability of the model predictions.

Scalability

To analyze larger fleets of batteries, instead of using leave-one-out methodologies, it may be more appropriate to de-

velop generalized training methodologies. In this direction, one approach would be to cluster batteries that exhibit similar operation and degradation conditions. This strategy would enable capturing data diversity, which is a key property for ensemble strategies. Alternatively, a hierarchical modelling strategy may be adopted. This method involves a global model for overall battery behavior, supplemented by smaller models for specific groups, enabling precise adaptations without the need for separate models per battery. This strategy ensures scalability and flexibility in handling various battery operation and degradation conditions efficiently.

Noise Sensitivity

The proposed approach assumes a Gaussian noise to model the variability of the modeled process and measurements [cf. Eq. (4)]. To analyze the impact of Gaussian noise levels on prediction results, a sensitivity analysis has been performed. Figure 9 shows the obtained results.

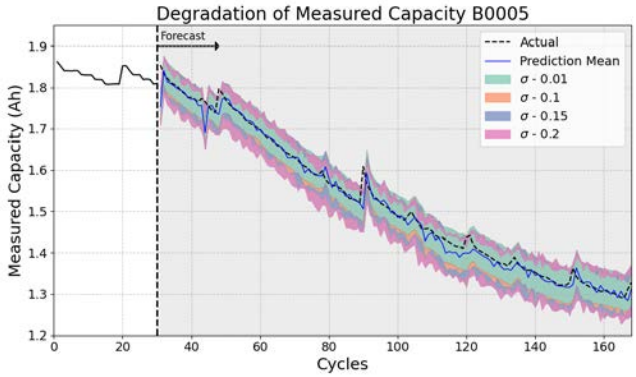


Figure 9. Impact of Gaussian Noise on Predictive Modeling of Battery Capacity Degradation.

Obtained results indicate that, when testing data diverges from training data, the epistemic uncertainty increases. The increase in Gaussian noise causes a greater deviation, and therefore, there is a significant rise in epistemic uncertainty. Analysing the model's behaviour in the presence of different types of uncertainty is crucial to evaluate the robustness of the model and determine if additional training stages are needed to enhance its reliability. Consequently, this research adopts a noise level of 0.1 as a trade-off decision between prediction accuracy and uncertainty.

Application Limits

Some of the adopted practices may limit the applicability of the proposed framework in real-world applications. The experimental setup, conducted in a controlled environment with specified load conditions, may not entirely replicate the diverse sources of uncertainty present in real-world applications. Such controlled conditions could potentially skew the understanding of uncertainty due to environmental and operational variabilities. Consequently, the predictive performance observed in this study may differ under less predictable conditions. In this direction, for controlled operation environments, the complexity of the proposed approach may be reduced. However, the proposed methodology complexity is designed to capture a wide range of uncertainties found in real operating systems.

6. Conclusion and Future Work

Batteries are key components in power and energy systems and ensuring a robust and reliable remaining useful life (RUL) prediction of batteries is crucial to develop accurate monitoring strategies, and build cost-effective solutions.

In this context, battery RUL prediction models generally focus on individual prediction models. They may be able to capture uncertainty associated with the battery ageing process, but the uncertainty modelling and capturing ability is also limited to the individual model. This research presents a probabilistic ensemble prognostics approach which combines Bayesian Convolutional Neural Network (BCNN) models in a probabilistic stacking strategy. The proposed framework leverages the probabilistic predictive information of individual BCNN models, which are integrated through a probabilistic stacking approach that calibrates between accuracy and robustness of probabilistic predictions.

The proposed approach has been tested on NASA's battery dataset. Obtained results show that the proposed probabilistic stacking approach improves accuracy and uncertainty of predictions with respect to other ensemble strategies and individual BCNN models.

This research study contributes towards understanding and predicting the capacity fade in Li-ion batteries. Namely, it highlights the role of probabilistic approaches and ensemble methods in modelling the uncertainties inherent in battery manufacturing and operation.

Looking forward, there are different opportunities to expand the scope and applicability of this work. On the one hand, the use of a larger battery dataset, which includes diverse environmental and operational conditions, would allow for a more comprehensive understanding of capacity fade across various scenarios. On the other hand, it may be possible to perform a more exhaustive comparative analysis of different fusion strategies, including Bayesian Model Averaging, Pseudo Bayesian Model Averaging, or Mixture Models. This comparative will provide further insights into the optimal approaches for integrating predictive models in the context of battery life prediction, enhancing both the accuracy and reliability of capacity fade forecasts.

Acknowledgenemts

This publication is part of the research projects KK-2023-00041, IT1451-22 and IT1676-22 funded by the Basque Gov-

ernment. J. I. Aizpurua is funded by Juan de la Cierva Incorporacion Fellowship, Spanish State Research Agency (grant No. IJC2019-039183-I).

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... Nahavandi, S. (2021, December). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. doi: 10.1016/j.inffus.2021.05.008

Bai, G., & Chandra, R. (2023, November). Gradient boosting Bayesian neural networks via Langevin MCMC. *Neurocomputing*, *558*, 126726. doi: 10.1016/j.neucom.2023.126726

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician*, *28*(1), 19–20.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In F. Bach & D. Blei (Eds.), *International conference on machine learning* (Vol. 37, pp. 1613–1622). Lille, France: PMLR.

Bosman, P. A., & Thierens, D. (2000). Negative log-likelihood and statistical hypothesis testing as the basis of model selection in ideas. In *Proceedings of the tenth dutch–netherlands conference on machine learning. tilburg university.*

Che, Y., Zheng, Y., Forest, F. E., Sui, X., Hu, X., & Teodorescu, R. (2024, January). Predictive health assessment for lithium-ion batteries with probabilistic degradation prediction and accelerating aging detection. *Reliability Engineering & System Safety*, *241*, 109603. doi: 10.1016/j.ress.2023.109603

Chung, Y., Char, I., Guo, H., Schneider, J., & Neiswanger, W. (2021). Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*.

Cobb, A. D., Himes, M. D., Soboczenski, F., Zorzan, S., O'Beirne, M. D., Baydin, A. G., ... Angerhausen, D. (2019, June). An Ensemble of Bayesian Neural Networks for Exoplanetary Atmospheric Retrieval. *The Astronomical Journal*, *158*(1), 33. doi: 10.3847/1538-3881/ab2390

Dai, H., Pollock, M., & Roberts, G. O. (2023, February). Bayesian fusion: Scalable unification of distributed statistical analyses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(1), 84–107. doi: 10.1093/jrsssb/qkac007

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., ... Saurous, R. A. (2017, November). *TensorFlow Distributions* (No. arXiv:1711.10604). arXiv.

Fan, Y., Olson, R., & Evans, J. P. (2017). A bayesian posterior predictive framework for weighting ensemble regional climate models. *Geoscientific Model Development*, *10*(6), 2321–2332.

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. doi: 10.1175/MWR2904.1

Hadigol, M., Maute, K., & Doostan, A. (2015). On uncertainty quantification of lithium-ion batteries: Application to an lic6/licoo2 cell. *Journal of Power Sources*, *300*, 507–524.

Hodson, T. O. (2022). Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, *2022*, 1–10.

Jung, Y., Jo, H., Choo, J., & Lee, I. (2022, June). Statistical model calibration and design optimization under aleatory and epistemic uncertainty. *Reliability Engineering & System Safety*, *222*, 108428. doi: 10.1016/j.ress.2022.108428

Kuleshov, V., Fenner, N., & Ermon, S. (2018, July). Accurate uncertainties for deep learning using calibrated regression. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2796–2804). PMLR.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641–1650.

Lee, G., Kwon, D., & Lee, C. (2023). A convolutional neural network model for SOH estimation of Li-ion batteries with physical interpretability. *Mechanical Systems and Signal Processing*, *188*, 110004. doi: 10.1016/j.ymssp.2022.110004

Liu, Y., Sun, J., Shang, Y., Zhang, X., Ren, S., & Wang, D. (2023, May). A novel remaining useful life prediction method for lithium-ion battery based on long short-term memory network optimized by improved sparrow search algorithm. *Journal of Energy Storage*, *61*, 106645. doi: 10.1016/j.est.2023.106645

Nam, G., Yoon, J., Lee, Y., & Lee, J. (2021). Diversity matters when learning from ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 8367–8377). Curran Associates, Inc.

Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., ... Hu, C. (2023). Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, *205*, 110796. doi: 10.1016/j.ymssp.2023.110796

Saha, B., & Goebel, K. (2007). Nasa ames prognostics data repository. *NASA Ames, Moffett Field, CA, USA*.

Toughzaoui, Y., Toosi, S. B., Chaoui, H., Louahlia, H.,

Petrone, R., Le Masson, S., & Gualous, H. (2022). State of health estimation and remaining useful life assessment of lithium-ion batteries: A comparative study. *Journal of Energy Storage*, *51*, 104520. doi: 10.1016/j.est.2022.104520

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., & Ulissi, Z. W. (2020, May). Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, *1*(2), 025006. doi: 10.1088/2632-2153/ab7e1a

Vanem, E., Salucci, C. B., Bakdi, A., & Alnes, Ø. Å. S. (2021, November). Data-driven state of health modelling—A review of state of the art and reflections on applications for maritime battery systems. *Journal of Energy Storage*, *43*, 103158. doi: 10.1016/j.est.2021.103158

Wang, C.-j., Zhu, Y.-l., Gao, F., Bu, X.-y., Chen, H.-s., Quan, T., ... Jiao, Q.-j. (2022). Internal short circuit and thermal runaway evolution mechanism of fresh and retired lithium-ion batteries with lifepo4 cathode during overcharge. *Applied Energy*, *328*, 120224.

Wei, Y., & Wu, D. (2023, February). Prediction of state of health and remaining useful life of lithium-ion battery using graph convolutional network with dual attention mechanisms. *Reliability Engineering & System Safety*, *230*, 108947. doi: 10.1016/j.ress.2022.108947

Yang, Y., Chen, S., Chen, T., & Huang, L. (2023). State of health assessment of lithium-ion batteries based on deep gaussian process regression considering heterogeneous features. *Journal of Energy Storage*, *61*, 106797.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018, September). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3). doi: 10.1214/17-BA1091

Zamo, M., & Naveau, P. (2018, February). Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences*, *50*(2), 209–234. doi: 10.1007/s11004-017-9709-7

Zhang, S., Liu, Z., & Su, H. (2022). A bayesian mixture neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Transportation Electrification*, *8*(4), 4708–4721. doi: 10.1109/TTE.2022.3161140

Zhang, Y., Zhang, M., Liu, C., Feng, Z., & Xu, Y. (2024, February). Reliability enhancement of state of health assessment model of lithium-ion battery considering the uncertainty with quantile distribution of deep features. *Reliability Engineering & System Safety*, 110002. doi: 10.1016/j.ress.2024.110002

Zhao, H., Chen, Z., Shu, X., Shen, J., Lei, Z., & Zhang, Y. (2023). State of health estimation for lithium-ion batteries based on hybrid attention and deep learning. *Reliability Engineering & System Safety*, *232*, 109066. doi: 10.1016/j.ress.2022.109066

Zhao, X., Wang, Z., Li, E., & Miao, H. (2024, January). Investigation into Impedance Measurements for Rapid Capacity Estimation of Lithium-ion Batteries in Electric Vehicles. *Journal of Dynamics, Monitoring and Diagnostics*. doi: 10.37965/jdmd.2024.475

Towards Efficient Operation and Maintenance of Wind Farms: Leveraging AI for Minimizing Human Error

Arvind Keprate[1,2] Stine. S. Kilskar[3] and Pete Andrews[4]

[1]*GrønnMet – Green Energy Lab, Department of Mechanical, Electrical and Chemical Engineering, Oslo Metropolitan University, Oslo, Norway*

[2]*AI Lab, Department of Computer Science, Oslo Metropolitan University, Oslo, Norway*

[3]*Department of Software Engineering, Safety and Security, SINTEF Digital, Trondheim, Norway*

[4]*EchoBolt, Alcester, UK*

arvind.keprate@oslomet.no

ABSTRACT

To effectively compete with other renewable energy sources, there remains a critical need to further decrease the Levelized Cost of Energy of Wind Farms (WFs). A promising way to achieve this objective is by minimizing the downtime of wind turbines (WTs) through effective Inspection and Maintenance (I&M) activities. Conventionally, I&M plans have predominantly relied on CM/SCADA data obtained from the physical components of turbines, with data analytics and machine learning (ML) techniques being employed to predict their performance and maintenance needs. However, statistics indicate that nearly 40% of WT failures can be traced back to HFs. These include aspects such as skills, knowledge, communication, and even the broader organizational culture. This paper delves into the importance of integrating HFs in the I&M of WFs to optimize turbine performance, enhance safety, and reduce downtime.

Firstly, we briefly discussed various Human Reliability Analysis (HRA) methods with special emphasis on Performance Shape Factors (PSFs). We then identify key human factors (HFs) that are vital for performing O&M tasks. For this, we have prepared a questionnaire to get qualitative input from technicians and also done a thorough literature review. E.g., some of the HFs that stand out include the ergonomics of tools and workspace designs tailored to technicians' needs, the cognitive load placed on operators during system monitoring and diagnostics, continuous training to handle evolving challenges, effective communication channels, and safety protocols designed with human behavior in mind. We then propose a novel framework for developing a computer vision-based recommendation system that can guide the technicians to perform the maintenance effectively thus minimizing the HE.

1. INTRODUCTION

The wind industry, driven by a commitment to green energy generation, is at the forefront of research, technological innovation, efficiency gains, and cost reductions. With turbine sizes and capacity factors having tripled, there has been a monumental shift in the wind energy sector. Since 1990, generation costs have been reduced by 65% (KPMG, (2019)), underscoring the industry's dedication to developing sustainable and economical energy solutions for the future. For instance, breakthroughs in blade design and materials, backed by rigorous research, enable turbines to harness wind more proficiently, yielding higher energy outputs even in suboptimal wind conditions (Asim, T., Islam, S., Hemmati, A., & Khalid, M. (2022)). The adoption of various Prognostics and Health Management (PHM) technologies and predictive analytics has further improved the operation and maintenance (O&M) of (WFs), curtailing downtime and driving costs even lower (Haghshenas, A., Hasan, A., Osen, O., & Mikalsen, E. T. (2023)).

Rinaldi et al. (Rinaldi, G., Thies, P. R., & Johanning, L. (2021)) performed an exhaustive survey of the latest strategies governing the O&M planning and CM of OWFs. Their review delves into the benefits and limitations of current practices and looks ahead to emerging trends in robotics, AI, and data analytics. Key opportunities highlighted include the integration of diverse data sources to refine O&M strategies, precise inventory management, detailed uncertainty modeling, the urgent need for

standardized open data frameworks, and the development of essential reference software. In a related study, McMorland et al. (McMorland, J., Flannigan, C., Carroll, J., Collu, M., McMillan, D., Leithead, W., & Coraddu, A. (2022)) highlighted the significance of various factors in O&M modeling for OWFs, including weather dynamics, failure, and degradation patterns, vessel logistics, cost estimation, and maintenance tactics. Besnard et al. (Besnard, F., Patriksson, M., Stromberg, A. B., Wojciechowski, A., & Bertling, L. (2009)) introduced the 'opportunistic maintenance' concept for OWFs, which entails the fusion of multiple planned corrective and preventive maintenance tasks, either within a similar timeframe or even during a single visit. By capitalizing on wind forecasts and synchronizing corrective maintenance with periods of low power generation or unexpected failures, this approach has proven to yield a 43% reduction in preventive maintenance expenses (Fast, S., Mabee, W., Baxter, J., Christidis, T., Driver, L., Hill, S., McMurtry, J., & Tomkow, M. (2016)) However, as currently practiced, the PHM approach uses only machine-related quantitative data available from CM/SCADA systems to predict and manage the performance and maintenance needs of WFs. The biggest drawback of the overreliance on machine-related (MR) data is its inability to capture the full spectrum of operating conditions under which WFs function. A frequently undervalued metric in this context is human-related data, which offers additional insights into the system environment (Kiassat, A.C., (2013)).

Human technicians/operators are an essential part of the daily O&M activities of the WFs. It is highly probable that Human Error (HE), in one form or another, might infiltrate the design, manufacturing, operation, and maintenance phases of WFs. Morag et al. (Morag, I. et al. (2018)), identified the most common HE during a maintenance activity described in Table 1.

The HE may go unnoticed due to various reasons and can result in catastrophic accidents leading to severe consequences for the environment, society, and business. Statistics indicate, HE as one of the major factors for accidents across various sectors as shown in Figure 1. For instance, the infamous disasters within the oil and gas sector namely, the Piper Alpha and the BP Deepwater Horizon blowout occurred due to human and operational flaws. Likewise, the accident investigations of multiple aircraft crashes (such as of a Boeing 707-321C in 1977; Boeing 747-200, in 1992; and Airbus 380-842 Qantas Flight 32 in 2010) also point towards technical failures, HFs, and regulatory shortcomings as failure causes (Mathavara, K., & Ramachandran, G. (2022)). These statistics serve as a reminder that, while the hardware aspect is undoubtedly important, the human dimension also has a significant influence on the overall health and performance of the system.

Table 1. Most common causes of Human Errors

(Morag, I. et al. (2018))

HE Type	**Description**
Communication	Misunderstandings among technicians and operators, often stemming from inadequate leadership and management.
Fatigue	Tiredness due to overwork or working in enclosed environments.
Tools and equipment	Improper use of tools and equipment can augment risks and compromise worker safety. Additionally, the lack of proper tools may increase HE as workers resort to using unsuitable machinery for specific tasks.
Skills and expertise	The risk of HE increases in non-routine tasks that demand specific knowledge when workers assigned are unfamiliar with the activities.
Bad procedures	HE often arises from poor information and the lack of standardized procedures.
Documentation	Poor documentation handling can increase HE due to its impact on task performance and understanding of required work.
Procedure's usage	Lengthy procedures often lead workers to adopt informal methods and rely on personal experience to complete tasks.
Time pressures	Overtime and overwork often lead to more mistakes by workers, as they resort to shortcuts and simpler work methods.
Tool control and housekeeping	It concerns tracking the equipment used or removed from machinery.

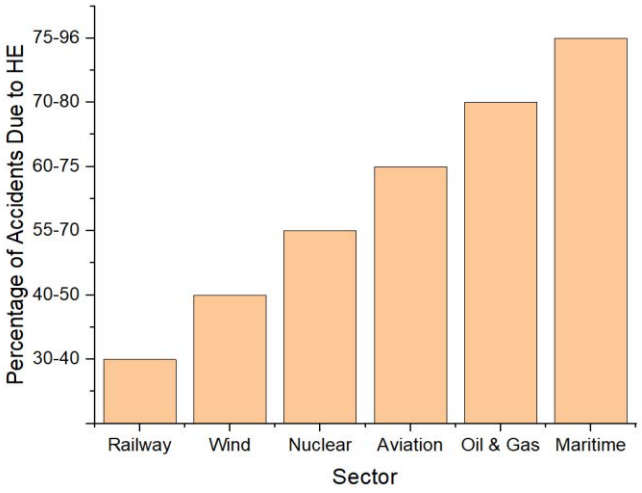


Figure 1. Accident percentage due to HE across various sectors

In this paper, the authors have highlighted the importance of integrating HFs within O&M of WFs. Firstly, we have briefly discussed various Human Reliability Analysis (HRA) methods with special emphasis on Performance Shape Factors (PSFs). We then discuss two scenarios of performance maintenance in the yaw deck and the nacelle of a typical WT. Thereafter we propose a framework for developing a computer vision-based recommendation system that can guide the technicians to perform the maintenance effectively thus minimizing the HE. We also propose the use of an eye-tracking device to measure the stress level of technicians.

2. HUMAN RELIABILITY ANALYSIS (HRA)

2.1. General

The origin of HRA is in probabilistic risk assessment (PRA), a discipline initially developed for understanding and quantifying the risks of serious accidents within the nuclear industry. HRA provides methods and tools for analyzing and assessing risks caused by operator's actions on a technical system, thus evaluating to operator's contribution to system reliability. The first fully developed HRA methods date back to the 1970s when systematic tools for analysis of the operator's contribution to risk were applied in the nuclear industry. There are now several HRA methods available for the nuclear sector, with some being adapted to other industries such as oil and gas, chemical, and aviation. Figure 2 illustrates the steps of a generic HRA process.

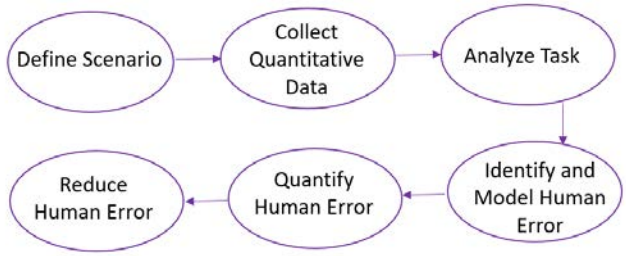


Figure 2. Generic HRA Process

2.2. HRA Methods

It is common to distinguish between first and second-generation HRA methods (Swain, A.D. (1990), Dougherty, E.M. (1990)). The list of first-generation methods is extensive and includes amongst others Technique for Human Error Rate Prediction (THERP) (Swain, A.D., Guttman, H.E. (1983)), the Human Cognitive Reliability method (HCR) (Hannaman, G.W., Spurgin, A.J., Lukic, Y.D. (1984)), the Human Error Assessment and Reduction Technique (HEART) (Williams, J.C. (1985)), Accident Sequence Evaluation Program (ASEP) (Swain, A.D. (1987), and Standardized Plant Analysis Risk – Human (SPAR-H) reliability analysis (Gertman, D., Blackan, H.S., Marble, J., Byers, J., Haney, L.N., Smith, C. (2005)).

Hollnagel (Hollnagel, E. (1998)), and Kim (Kim, I.S. (2001)) provide the following list of notable characteristics of first-generation methods: 1) Assumption that human reliability is similarly describable as hardware reliability. 2) HRA being limited to only the human actions that are included in the PSA event trees. 3) Binary representation of human action as either success or failure to carry out a given task. 4) Dichotomy of errors of omission (failure to perform an action) and errors of commission (unintended or unplanned action). 5) Focus on phenomenological aspects of human actions. 6) Little concern about the cognitive aspects of human actions. 7) Emphasis on quantification of human errors. 8) Indirect treatment of context, as the way in which PSFs exert their effect on performance is not described.

Second-generation HRA methods were developed based on cognitive architectures to unveil the causes of errors from a behavioral perspective; thus, solving the main deficiency of the first generation. Two basic requirements proposed by Hollnagel (Hollnagel, E. (1998)) are that second-generation approach "uses enhanced PSA event trees and that it extends the traditional description or error modes beyond the binary categorization of success-failure and omission-commission" (p.151). He further stresses the need for a more realistic type of operator model, as the approach must be explicit about the way in which performance conditions affect performance. Most authors critiquing first-generation HRA methods agree on the necessity of incorporating a cognitive model into HRA "that would enable a better understanding of human error mechanisms that were well described by Reason (Reason, J. (1990))". A Technique for Human Event Analysis (ATHEANA) (Cooper, S.E., Ramey-Smith, A.M., Wreathall, J., Parry, G.W. (1996)) and Cognitive Reliability and Error Analysis Method (CREAM) (Hollnagel, E. (1998)) are examples of well-known and widely utilized second-generation techniques. CREAM uses the contextual control model (COCOM) and provides a determination of the reliability of a person's performance based on an error taxonomy that contains both error modes and error causes.

Although addressing the main issue of first-generation HRA methods, one of the highlighted weaknesses of second-generation methods is that they do not provide sufficient consideration of the mutual influences between PSFs (De Ambroggi, M. (2011)). According to Griffith and Mahadevan (Griffith, C.D., Mahadevan, S. (2011)) the main sources of deficiencies in HRA methods include: "1) lack of empirical data for model development and validation, 2) lack of inclusion of human cognition (i.e., need for better human behavior modeling, 3) large variability in implementation (i.e., HRA parameters are different depending on the method used), and 4) heavy reliance on expert judgment in selecting PSFs, and use of these PSFs to obtain the HEP in human reliability analysis" (p. 1444).

HRA experts have more recently begun to look at potential improvements to existing methods. As an example, the

HEART method has been used as a basis for domain-specific approaches such as Nuclear Action Reliability Assessment (NARA) (Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G., Umbers, I. (2004)), Controller Action Reliability Assessment (CARA) (Kirwan, B., Gibson, H. (2008)), Railway Action Reliability Assessment (RARA) (Gibson, W.H., Mills, A.M., Smith, S., Kirwan, B.K. (2013)) and Shipboard Operations Human Reliability (SOHRA) (Akyuz, E., Celik, M., Cebi, S. (2016)). Another example is a more recent article by He et al. (He, Y., Kuai, N.-S., Deng, L.-M., He, X-Y. (2021)), which builds on CREAM by adding Human Inherent Factors (HIFs) such as anti-fatigue ability, concentration ability, reaction ability, and personality traits.

In 2006, NASA Office of Safety and Mission Assurance (OSMA) published a technical report evaluating 14 HRA methods against a list of 17 attributes to highlight methods that are considered suitable for use in risk and reliability studies of NASA space systems and missions. The evaluation resulted in the selection of four methods: THERP, CREAM, NARA, and SPAR-H. The list of attributes used to compare the methods included: Developmental Context, Screening, Task Decomposition, PSF List and Causal Model, Coverage, HEP Calculation Procedure, Error-Specific HEPs, Task Dependencies and Recovery, HEP Uncertainty Bounds, Level of Knowledge Required, Validation, Reproducibility, Sensitivity, Experience Base, Resource Requirements, Cost and Availability, as well as Suitability for NASA Applications (Chandler, F., Chang, Y., Mosleh, A., Marble, J., Boring, R., Gethman, D. (2006)). Consideration of several of these attributes is essential when evaluating existing HRA methods for use in the context of O&M of WFs.

2.3. Human Factors

Our definition of HFs is from IEA: "Human Factors is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and other methods to design to optimize human well-being and overall system performance". HFs can be used either in accident investigations, or they can be used to enhance the performance of the technicians.

The aims of using HF in general and in accident investigations are to:

(1) Improve safety (i.e., reducing the risk of injury and death);

(2) Improve performance in safety-critical situations (i.e., increase quality, productivity, and efficiency);

(3) Support satisfaction/usability (i.e., increasing acceptance, comfort, and well-being).

The details of how to use HFs for accident investigation are well documented in the literature, however, in this paper, we shall focus more on the identification of the HFs (in particular PSFs) that can be managed such that the I&M activities are performed efficiently within given time with minimal HE.

2.4. Performance Shape Factors (PSFs) for OWFs

PSFs or Performance Influencing Factors (PIFs) are defined by the Health Safety and Executive (HSE) as “characteristics of the job (e.g. the working environment); the individual (physical capability to do the work), and the organization (e.g. time pressure) that influence human performance” (HSE RR01 (2002))

Relevant PSFs for OWFs include environmental conditions (e.g., high winds, rough seas, weather variability), ergonomic challenges (working at heights, confined spaces, awkward postures), organizational aspects (training, work culture, resource availability), technical and mechanical complexity, accessibility and logistics due to remote locations, communication and coordination for emergency response, and the use of specialized tools and predictive maintenance technologies. On a more personal level, psychological stressors such as time pressure and distractions, as well as physiological factors like fatigue and hunger, can impact inspection and maintenance quality and error rates, especially in confined spaces like nacelles and hubs.

Acknowledging PSFs and their impact on operational outcomes is essential for ensuring the safety, efficiency, and reliability of OWF’s O&M. For instance, the performance of technicians can significantly drop on a wet and windy day compared to more favorable weather conditions, increasing the risk of human error and injuries. Similarly, an overloaded technician may overlook early signs of wear, potentially causing unforeseen equipment failures. Additionally, a company that prioritizes proactive maintenance is likely to emphasize regular training, which can lead to fewer operational errors.

The I&M activities and corresponding PSFs differ depending on the location within the WTs. For example, tasks on the yaw deck, such as brake maintenance and friction pad replacement, present unique challenges. These include transporting items using the nacelle crane or manually from inside the tower. Operations in this area entail inspecting the deck, handling moving parts, setting up the workspace, conducting maintenance, and cleaning up (G+ Global Offshore Wind Health & Safety Organization, (2021)). Challenges specific to the yaw deck include difficult access, particularly through ladder hatches in older turbines, constrained working space, and the physical strain of maneuvering heavy items. These conditions require technicians to employ specialized tools and assume strenuous postures, which can adversely affect their well-being (G+ Global Offshore Wind Health & Safety Organization, (2021)).

Most service and maintenance tasks in WTs, such as routine inspections and part replacements, are carried out in the confined spaces of the nacelle and blade hub. Although newer, larger wind turbines provide a bit more space and improved accessibility, the areas remain constrained, frequently cluttered, and occasionally slippery due to oil spills. These conditions make it difficult to move safely and operate tools efficiently (G+ Global Offshore Wind Health & Safety Organization, (2021)). In the hub, accessing components like blade root bolts forces technicians into uncomfortable positions, compounded by the presence of grease and cramped, angled spaces. This increases the likelihood of injuries, equipment mishandling, and errors. More details regarding PSFs for working on OWT can be found in (G+ Global Offshore Wind Health & Safety Organization, (2021)).

A questionnaire was designed to collect feedback from technicians, the results of which are presented in Figure 3 (in the Appendix). The questionnaire's link is provided in (Questionnaire, (2024)). The responses indicate a consensus among technicians on most questions. For instance, regarding ergonomic challenges highlighted in question 3, one technician mentioned, "*Wind turbines are often not ergonomically designed, lacking laydown areas for bags and equipment, leading to obstacles and potential hazards. Restricted access, working in areas with significant grease or oil, and maintaining a clean environment pose substantial challenges.*" Another respondent highlighted the absence of adequate sanitary facilities for women on WTs.

A detailed analysis of the survey results suggests that conducting I&M activities on WTs is an exceptionally challenging task, which significantly increases the likelihood of HE. Moreover, the lack of real-time supervision at inspection sites reduces the opportunities to correct such errors. Consequently, the following section introduces a novel framework for a Computer Vision supervisory agent designed to monitor technicians during inspections and capable of raising an alarm if there is a risk of HE

3. COMPUTER VISION-BASED RECOMMENDATION AGENT

The steps involved in the framework that integrates multi-modal inputs like videos and images consist of the following steps:

1. Data Collection: Using high-resolution cameras, we will gather a comprehensive dataset of videos and images capturing expert technicians performing WT inspections.
2. Data Preprocessing: We will apply techniques like frame extraction, noise reduction, and image stabilization to the recorded videos and images to prepare the data for analysis. Next, we will manually annotate them with labels indicating correct and incorrect actions, focusing on key inspection points and common errors. Finally, we will augment the data using techniques such as rotation, scaling, and mirroring to increase the dataset's robustness against variations in real-world scenarios.
3. Model Development: We will use convolutional neural networks (CNNs) to extract features from images and video frames, and employ Long Short-Term Memory (LSTM) networks to analyze temporal dependencies in video data. We will then implement a fusion technique to effectively integrate features from different modalities, capturing a comprehensive profile of inspection activities. Lastly, we will develop a classification system using machine learning to distinguish between correct and incorrect inspection behaviors based on the labeled data.
4. Real-Time Monitoring System: We will install a monitoring device at strategic locations around the wind turbine. Each device will be equipped with a high-resolution camera and a speaker system. The camera will continuously capture video of the technician’s activities, allowing the system to visually monitor the inspection process from multiple angles. We will use edge computing devices integrated within the monitoring systems to process the data in real-time, significantly reducing latency and ensuring that any deviations or anomalies are promptly detected. The speaker will provide immediate audio feedback and recommendations to the technician based on the real-time analysis, including alerts about potential errors, reminders of inspection steps, or safety warnings.
5. Feedback Loop: We will integrate a feedback system where the model learns from new inspection videos over time, adapting to new techniques and evolving standards in turbine maintenance. We will regularly evaluate the system’s accuracy and reliability in detecting deviations and making iterative improvements based on real-world performance and feedback from technicians and supervisors. Furthermore, the technicians will also be able to interact with the system using voice commands. They will be able to respond to the audio cues by confirming receipt of messages or asking for further clarification. They will also be able to report issues, fetch information, or even tag certain observations without having to stop their work or remove their gloves, which can be particularly useful in harsh weather conditions.

The deployment of such a framework has the potential to lower the HE significantly within WT maintenance and it also aligns with the broader goals of the wind industry to reduce costs and improve the reliability and efficiency of green energy production. As the industry continues to evolve, the continuous refinement and adoption of such integrated frameworks will be essential for sustaining

growth and ensuring the safety and well-being of the human technicians at the heart of these operations.

4. CONCLUSION

This paper laid out the critical importance of integrating HFs into the O&M of WFs, with a particular focus on the potential to enhance safety and efficiency through advanced technologies and methodologies. We discussed various approaches that have been used in the past for performing HRA to estimate HEP. The important PSFs for maintenance activity on WFs, include environmental conditions, ergonomic challenges, organizational aspects, accessibility and logistics due to remote locations, communication and coordination for emergency response, the use of specialized tools, and psychological stressors. A questionnaire was designed to collect feedback on PSFs from WT technicians. For example, all the technicians agreed that the awkward positions required for accessing components like blade root bolts not only increase the risk of injury but also elevate the likelihood of mishandling equipment and making errors.

To address these issues, we proposed a computer vision-based supervisory agent capable of real-time monitoring. This system, which utilizes multi-modal inputs from high-resolution cameras and provides audio feedback, represents a significant leap forward in reducing HE. By continuously capturing and analyzing the technician's actions, the system offers corrective feedback and actionable recommendations, thereby ensuring adherence to best practices and enhancing overall safety.

REFERENCES

1. KPMG, 2019. https://assets.kpmg.com/content/dam/kpmg/dk/pdf/DK-2019/11/The-socioeconomic-impacts-of-wind-energy_compressed.pdf
2. Asim, T., Islam, S., Hemmati, A., & Khalid, M. (2022, January 14). A Review of Recent Advancements in Offshore Wind Turbine Technology. Energies, 15(2), 579.
3. Haghshenas, A., Hasan, A., Osen, O., & Mikalsen, E. T. (2023, January 25). Predictive digital twin for offshore wind farms. Energy Informatics, 6(1).
4. Rinaldi, G., Thies, P. R., & Johanning, L. (2021, April 27). Current Status and Future Trends in the Operation and Maintenance of Offshore Wind Turbines: A Review. Energies, 14(9), 2484.
5. McMorland, J., Flannigan, C., Carroll, J., Collu, M., McMillan, D., Leithead, W., & Coraddu, A. (2022, September). A review of operations and maintenance modelling with considerations for novel wind turbine concepts. Renewable and Sustainable Energy Reviews, 165, 112581.
6. Besnard, F., Patriksson, M., Stromberg, A. B., Wojciechowski, A., & Bertling, L. (2009, June). An optimization framework for opportunistic maintenance of offshore wind power system. 2009 IEEE Bucharest PowerTech.
7. Fast, S., Mabee, W., Baxter, J., Christidis, T., Driver, L., Hill, S., McMurtry, J., & Tomkow, M. (2016, January 25). Lessons learned from Ontario wind energy disputes. Nature Energy; Nature Portfolio.
8. Kiassat, A.C., 2013. System Performance Analysis Considering Human-related Factors. PhD Thesis. University of Toronto.
9. Morag, I. et al. (2018) 'Identifying the causes of human error in maintenance work in developing countries', International Journal of Industrial Ergonomics, 68, pp. 222–230.
10. Mathavara, K., & Ramachandran, G. (2022). Role of Human Factors in Preventing Aviation Accidents: An Insight. IntechOpen. doi: 10.5772/intechopen.106899.
11. Swain, A.D. (1990). Human reliability analysis: need, status, trends and limitations. *Reliability Engineering and System Safety* 29(3), 301-313.
12. Dougherty, E.M. (1990). Human reliability analysis – where should thou turn? *Reliability Engineering and System Safety* 29(3), 283-299.
13. Swain, A.D., Guttman, H.E. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications. Final report. NUREG/CR-1278. Washington, DC: US Nuclear Regulatory Commission.
14. Hannaman, G.W., Spurgin, A.J., Lukic, Y.D. (1984). Human Cognitive Reliability Model for PRA Analysis. Draft Report NUS-4531, EPRI Project RP2170-3. Electric Power and Research Institute, Palo Alto, CA.
15. Williams, J.C. (1985), HEART A proposed method for achieving high reliability in process operation by means of human factors engineering technology, in *Proceeding of a symposium on the achievement of reliability in operating plant, Safety and Reliability Society*, 16 September 1985.
16. Swain, A.D. (1987). Evaluation of Human Reliability on the Basis of Operational Experience, in Economics and Social Science. The Munich Technical University.
17. Gertman, D., Blackan, H.S., Marble, J., Byers, J., Haney, L.N., Smith, C. (2005). The SPAR-H Human Reliability Analysis Method. U.S. Nuclear Regulatory Commission. NUREG/CR-6883, Washington DC.
18. Hollnagel, E. (1998). Cognitive Reliability and Error Analysis Method CREAM. 1. Ed., Elsevier.
19. Kim, I.S. (2001). Human Reliability analysis in the man-machine interface design review. *Annals of Nuclear Energy* 28(11), 1069-1081.

20. Reason, J. (1990). Human Error. Cambridge University Press, Cabridge, UK.
21. Cooper, S.E., Ramey-Smith, A.M., Wreathall, J., Parry, G.W. (1996). A technique for human error analysis (ATHEANA). Technical basis and methodology description. USNRC; 1996. No. Nureg/CR-6350.
22. De Ambroggi, M. (2011). Modelling and assessment of dependent performance shaping factors through Analytic Network Process. *Reliability Engineering and System Safety* 96(7), 849-860.
23. Griffith, C.D., Mahadevan, S. (2011). Inclusion of fatigue effects in human reliability analysis. *Reliability Engineering and System Safety* 96(11), 1437-1447.
24. Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G., Umbers, I. (2004). Nuclear action reliability assessment (NARA): a data-based HRA tool. In: *Probabilistic safety assessment and management.* Springer, p. 1206-1211.
25. Kirwan, B., Gibson, H. (2008). CARA: a human reliability assessment tool for air traffic safety management – technical basis and preliminary architecture. In: *The safety of systems*. Springer, p. 197-214.
26. Gibson, W.H., Mills, A.M., Smith, S., Kirwan, B.K. (2013). Railway action reliability assessment, a railway-specific approach to human error quantification. In: *Proceedings of the Australian system safety conference*, 7 p.
27. Akyuz, E., Celik, M., Cebi, S. (2016). A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique. *Safety Science* 87, 108-122.
28. He, Y., Kuai, N.-S., Deng, L.-M., He, X-Y. (2021). A method for assessing Human Error Probability through physiological and psychological factors tests based on CREAM and its applications. *Reliability Engineering and System Safety* 215 (2021) 107884, 12 p.
29. Chandler, F., Chang, Y., Mosleh, A., Marble, J., Boring, R., Gethman, D. (2006). Human Reliability Analysis Methods: Selection Guidance for NASA. NASA Office of Safety and Mission Assurance, Washington, DC (2006), 123 p.
30. HSE RR01 (2002). Human factors integration: Implementation in the onshore and offshore industries.
31. G+ Global Offshore Wind Health & Safety Organization, 2021 incident report. Energy Institute, UK.
32. Questionnaire,2024.https://forms.office.com/Pages/ShareFormPage.aspx?id=Eh_I_oZiUEWJEfRG_Nr6HwUVrL3FuU9GkTQNYRKR5FURTFNWDgxSEJSQ0U5TE5IVlo3TDRRTVZPSy4u&sharetoken=4iZz38x8ISItUPxLNwXM

BIOGRAPHIES

Arvind Keprate received his B. Tech in Mechanical Engineering (2007) from Himachal Pradesh University, M.Sc. in Marine & Subsea Technology (2014), and PhD (2017), in Offshore Engineering from the University of Stavanger, Norway. He is currently a Professor at Oslo Metropolitan University where he teaches Design related courses to Mechanical Engineering students. Besides this, he also teaches Machine Learning, Probability & Statistics at Kristiania University College in Oslo. He has been a visiting researcher at the Prognostics Center of Excellence, NASA Ames Research Center, USA. Currently, his research is focused on Digital Twins and PHM of complex Socio-Ecological-Technical Energy Systems such as Wind Farms.

Stine Skaufel Kilskar has an M.Sc. in Industrial Economics and Technology Management (2014) from the Norwegian University of Science and Technology, Norway, with a focus on strategic change management. She is currently a Research Scientist at SINTEF Digital in Trondheim. She has ten years' experience of working in safety-related research projects within various industries, such as construction, maritime, oil & gas, and energy. The research is mainly focused on safety management and human factors.

Pete Andrews qualified from the University of Sheffield with a Masters degree in Aerospace Engineering in 2005. Working within the power industry for the last 19 years he has delivered operational, engineering and leadership roles across a broad range of power generation assets and technologies. Previously he delivered a number of roles within leading utilities including Commercial Manager supporting major asset divestments and managing offshore wind services and Plant Manager accountable for a large offshore wind farm. Recognizing that the pace of innovation in offshore wind operations and maintenance was lagging compared to the rate of development in other aspects of the sector he founded EchoBolt, an organisation dedicated to deploying advanced technologies to improve the management of structural integrity of wind turbines.

Appendix

Fig 3. Response of Questionnaire

1. **Environmental Conditions:** How often do adverse weather conditions (e.g., high winds, rain, sea states) affect your ability to safely perform maintenance tasks?

2. **Ergonomic Challenges:** Do you face any physical difficulties while working on turbines. How do these challenges affect your work?

3. **Ergonomics Challenges:** Please describe any physical difficulties you encounter while working on turbines (e.g., working in confined spaces, at heights, or in awkward positions)

Latest Responses

"Working in awkward positions for long time"

"very hard for women to be there in case of periods "

"Wind turbines are often not ergonomically designed. No laydown areas for ...

4. **Psychological Stressors:** How often do you find that time pressure or distractions affect your focus during maintenance tasks?

5. **Physiological Factors:** How frequently do fatigue or hunger impact your ability to perform maintenance and inspection tasks effectively?

6. **Organizational Support:** How adequate do you find the training and resources provided for dealing with the specific challenges of offshore wind turbine maintenance?

7. **Technical Complexity:** Rate the level of difficulty you face in understanding and applying the technical knowledge required for turbine maintenance.

8. **Accessibility and Logistics:** How do the logistics and accessibility of offshore wind farms impact your maintenance work, particularly in emergency situations?

9. **Communication and Coordination:** Evaluate the effectiveness of communication within your team and with management, especially for coordinating maintenance activities and responding to incidents.

10. **Safety and Emergency Procedures:** How confident are you in the safety protocols and emergency response plans in place for offshore wind turbine operations? Have you identified any areas for improvement?

Towards Physics-Informed PHM for Multi-component degradation (MCD) in complex systems

Atuahene Barimah[1], Octavian Niculita[2], Don McGlinchey[3], Andrew Cowell[4] and Billy Milligan[5]

[1,2,3,4] *Glasgow Caledonian University, Glasgow, G4 0BA, UK*
abarim300@gcu.ac.uk, octavian.niculita@gcu.ac.uk, d.mcglinchey@gcu.ac.uk,A.Cowell@gcu.ac.uk

[5]*Howden|Chart Industries, Glasgow, Renfrew, PA4 8XJ, UK*
billy.milligan@howden.com

ABSTRACT

This study seeks to address the challenge of limited degradation data in developing Fault Detection and Isolation (FDI) models for multi-component degradation (MCD) scenarios. Utilizing a small fraction (0.05%) of a previously utilized water distribution testbed dataset in a previous publication, a weighted ensemble hybrid approach is proposed and evaluated against more established modelling approaches used in the previous publication. The proposed approach combines heuristic approximation and Physics-Informed Neural Network (PINN) methods with a recurrent neural network (RNN) model to enhance diagnostic performance for predicting MCD scenarios. The hybrid model generally outperformed other algorithms when tested on an MCD dataset, demonstrating improved diagnostic accuracy in such scenarios. Future research aims to optimize ensemble weights based on model uncertainty, further enhancing diagnostic capabilities.

1. INTRODUCTION

Data has become the story of engineering design in recent times as the availability of system data provides insights into the dynamics of any complex system. This is particularly true for developing analytics in digital twin (DT) design for asset health management applications (Lu, Xie, Parlikad, & Schooling, 2020). Figure 1 shows the nexus between the analytics developed for PHM applications and a virtual representation of a physical asset highlighting how DTs can enable PHM applications. In exploring cost mitigation strategies, different maintenance data-driven models often rely on large amounts of data to train effectively (Maass, Parsons, Purao, Storey, & Woo, 2018). The more data is available, the better the model can learn patterns and relationships within the data, leading to more accurate predictions or insights (Barimah, Niculita, McGlinchey, & Cowell, 2023). This helps data-driven models generalise better to unseen data the more data is available (Duriez, Brunton, & Noack, 2017). This is particularly useful when it comes to asset health management where the availability of trainable degradation data is critical in the design and execution of Prognostics and Health Management (PHM) strategies for complex systems undergoing multi-component degradation scenarios.

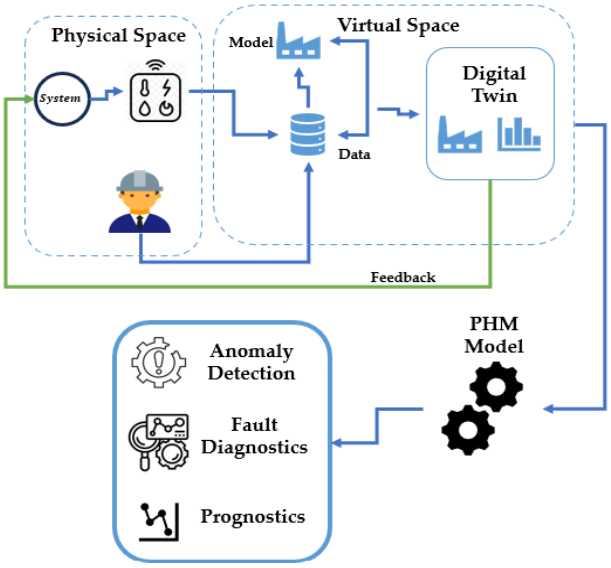


Figure 1. Relationship between DT and PHM applications.

However, obtaining asset degradation data can be expensive, time consuming, and often requires specialized equipment, sensors, or monitoring systems (Hu, Miao, Si, Pan, & Zio, 2022). Operators often rely on post-failure degradation data (Barimah, Niculita, McGlinchey, & Alkali, 2021) which enables the development of statistical-based techniques for

system-level anomaly detection. The statistical-based technique alone prevents an operator from isolating the sub-systems contributing to the anomaly in the larger system. The scenario becomes even more complex when two components of the same system degrade simultaneously (very often, at different rates). Simulated degradation data calibrated with the actual physical system (Higdon, Kennedy, Cavendish, Cafeo, & Ryne, 2004) can also be used to train predictive models. However, this approach becomes limited when MCD scenarios are being considered as degradation data from the different combinations of sub-systems undergoing degradation need to be simulated to generated the required data. For complex systems with a lot of sub-systems with different operating conditions, this approach becomes untenable.

To address the issue of limited data, several authors have suggested combining the insights given by data-driven models with some physical equations that govern the dynamics of a system. Physics-Informed Neural Networks (PINNs) integrate the physics of an asset into their training process, enforcing physical constraints alongside data-driven learning. PINNs can generalize well with limited data (Cai, Mao, Wang, Yin, & Karniadakis, 2021) and have applications in various fields (Huang & Wang, 2022), making them valuable for tackling complex, multi-physics problems (Bararnia & Esmaeilpour, 2022) by reducing computational costs and providing insights (Rizi & Abbas, 2023). The aim and objectives of this paper are presented in section 2 below.

2. OBJECTIVES OF STUDY

This paper aims to develop and benchmark an ensemble hybrid fault detection and isolation model for components (sub-systems) undergoing multi-component degradation (MCD) scenarios in a water distribution system.

- Identify a physical equation that represents the degradation severity level of either blockages or leakages in the system.
- Design a Fault Detection and Isolation (FDI) algorithm using a PINN-enabled Hybrid model for each component in the water distribution system.
- Train all FDI models on limited degradation data and test models on test multi-component degradation scenario data from the same system at different operating conditions.
- Identify areas of model improvement and potential research.

The paper is structured as follows: Section 3 covers the methodology. Sections 4 and 5 present and discuss the results of FDI model performance. Finally, the paper concludes with contributions and future research work.

3. METHODOLOGY

3.1. System Description

Data from the dynamic behaviour of a water distribution system undergoing multi-component degradation presented in Barimah et. al (2023) was used in this report. Figure 2 below shows the water distribution experimental testbed, where an external gear pump pumps water from a main supply tank. A variable speed drive (VSD) controls the rotational speed of the pump and the motor. The system also has five (5) direct proportional valves (DPV1 to DPV5) and a solenoid shut-off valve (SHV) that were included to support the emulation of deterioration phenomena affecting five different components in a controlled manner. Data is collected from five pressure transmitters (P1, P2, P3, P4, and P5), turbine flow meters (f1 and f2), and a laser sensor to gauge the pump's speed. Table 1 lists the control valves in the system's default operating states, their respective fault codes, and the fault emulation mechanism for each component on the testbed.

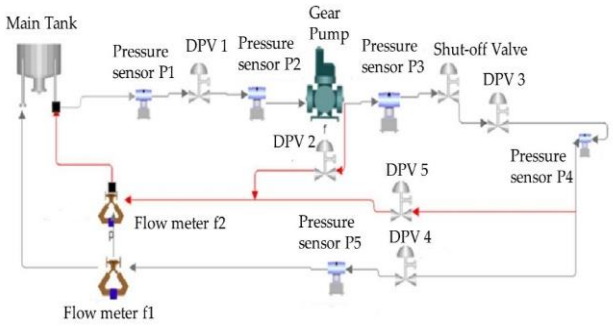


Figure 2. Water Distribution System Testbed Schematic (Barimah et. al 2023).

Table 1. Healthy condition operating state of the system's control valves and associated fault codes

Component/Fault Codes	Testbed Valves	Healthy State	Fault Emulation Mechanism
Filter/FC1	DPV 1	FO	DPV1 GC
Pump/FC2	DPV 2	FC	DPV2 GO
Valve/FC3	DPV 3	FO	DPV3 GC
Nozzle/FC4	DPV 4	FO	DPV4 GC
Pipe/FC5	DPV 5	FC	DPV5 GO

FO - Fully Open | FC - Fully Closed
GC - Gradually closing | GO - Gradually opening.

3.2. Process Data Capture

The degradation data used in the previous publication by Barimah et. al (2023) was recorded within four (4) weeks for healthy condition (HC), Single Component Degradation (SCD) and Multi-Component Degradation (MCD) scenarios between pump speeds of 700rpm and 950rpm in intervals of 50rpm. The SCD process data represents the degradation of individual components (See Table 2) with pressure and flow

measurements at $P_1, P_2 \ldots\ldots P_5$ and f_1, f_2 respectively. Data logging for each faulty condition scenario last between three (3) to four (4) minutes and also starts at least 10 minutes after the process reaches steady state conditions or when there is a step change in pump speed or a change in the failure condition scenario with each data file having different sample sizes. The degradation level of severity $(0 \leq S \leq 1)$ for each component on the testbed is determined by gradually closing or opening the respective direct proportional valve.

Table 2. Data capturing process with a sampling rate of 0.2s

Test period	4 consecutive weeks
Faulty Condition Scenarios (Total No of Tests)	FC0 – Healthy Condition (24) FC1–Clogged Filter (24) FC2–Degraded Pump (24) FC3–Blocked Valve (24) FC4–Blocked Nozzle (24) FC5–Leaking Pipe (24)
Pump Speed (rpm)	700/750/800/850/900/950

3.3. FDI Model Development

Sections 3.3.1 to 3.3.5 presents the process for developing physics informed fault detection and isolation (FDI) algorithms. Using limited training data, the paper also benchmarks the statistical process control (SPC), ensemble classification models and a recurrent neural network model presented by the author in a previous paper (see Figure 4) with the physics Informed FDI models presented in this paper. This is to determine the performance of FDI models in detecting multi-component degradation scenarios when limited degradation data is available for model training. The Statistical Process Control, the ensemble and neural network models were trained with full degradation dataset in Table 2 simultaneously, tagged as models $M_1, M_2 and\ M_3$ respectively and stored in a Fault Detection and Isolation model repository. The function $f_2(x)$ is then used to determine the proportion of accurate predictions of test degradation scenario data by $M_1, M_2 and\ M_3$.

A physics-informed Neural Network (PINN) enabled hybrid FDI algorithm is also developed in this report to detect multi component degradation scenarios. The hybrid model consists of a weighted average of a heuristic approximation model, a naïve recurrent neural network and a feedforward PINN model for each component in the water distribution system. The training of all models was done using randomly selected 0.05 % of the full degradation data from the original historical dataset used in Barimah et. al (2023) shown in Table 2. Figure 5 below shows the various cases in which the various randomly sampled degradation data can occur. Case A represents a scenario where part of the live process data from the system forms part of the distribution of the sampled random data. Case B is the non-ideal situation where the live process data is a subset of the sampled distribution while case C is the ideal case where the degradation data available is truly limited. The rationale for the random approach is to reduce the quantity and diversity of degradation data available for model training and development hence the limited nature. This is done to determine the impact of limited data conditions on the performance of FDI algorithms in MCD scenarios. The levels of severity are categorized into two (2) groups with below 0.21 defined as healthy and between 0.21 and 1.0 defined a faulty in both MCD and SCD scenarios (see Figure 3). The performance of the FDI algorithms is measured using the interval $(0 \leq Performance \leq 1)$ where 1 means the algorithm predicted correctly all the categorized severity states of the asset in operation while 0 means the algorithm failed to predict correctly any severity state of the asset.

3.3.1. Physics Informed Neural Network Model

As shown in Figure 2, five direct-acting proportional valves are used to emulate the dynamics of degradation patterns in the main components on the water distribution system. Equation (1) is used to determine the fluid flow through a valve where $f(V_0)$ is the function of valve opening with $0 \leq f(V_0) \leq 1$ as the interval for the valve opening and C_v is the valve coefficient (Knight, Russell, Sawalk & Yendell, 2013). Equations (2) & (3) are used to determine the level of severity $S(0,1)$ for blockages (Blocked Filter, degraded valve & Blocked Nozzle) and leakage (leaking pipe) respectively. For the pump, Eq. (4) is used to determine the severity level in a leaking pump degradation scenario for the gear pump on the testbed where N_v and N_m are the volumetric and mechanical efficiencies respectively. The maximum level of severity occurs when $S = 1$ with no fault condition being $S = 0$. Therefore, the interval of degradation for each component on the testbed is $0 \leq S \leq 1$ (see Figure 3).

$$flow = C_v f(V_0)\sqrt{\frac{\Delta P}{SG}} \quad (1)$$

$$S(0,1) = 1 - f(V_0) \quad (2)$$

$$S(0,1) = f(V_0) \quad (3)$$

$$S(0,1) = 1 - (N_m N_v) \quad (4)$$

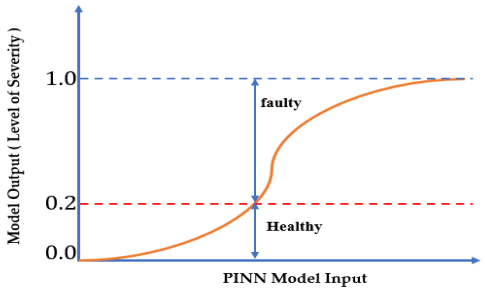


Figure 3. Change in component degradation severity level.

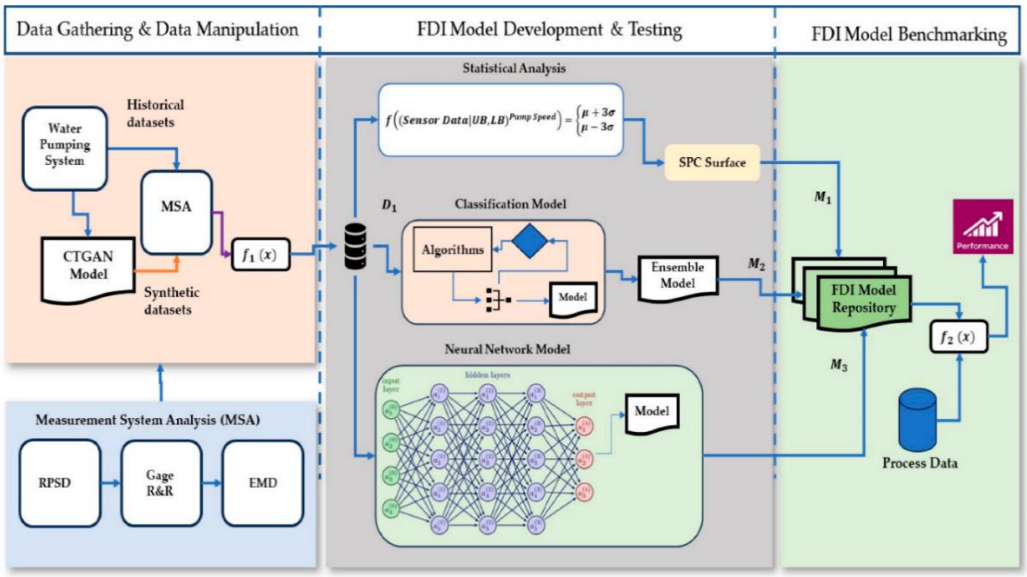


Figure 4. Proposed process for benchmarking the FDI algorithms (Barimah et. al 2023).

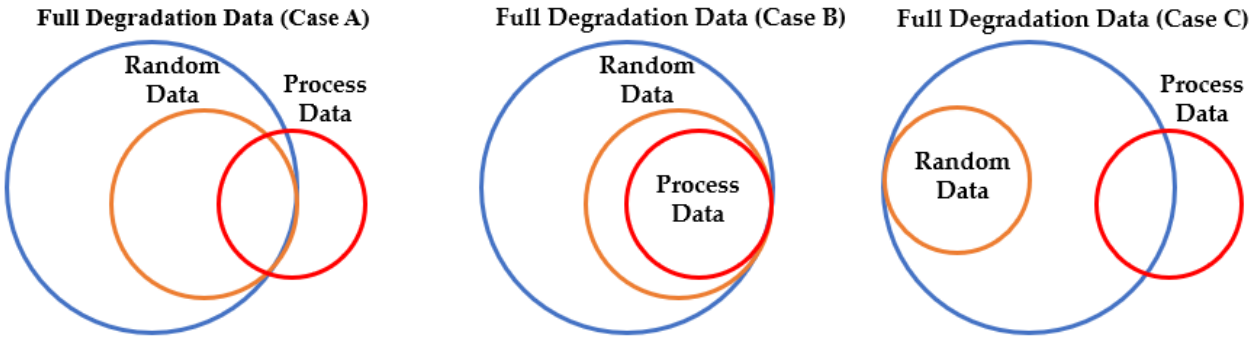


Figure 5. Random sampling of full degradation dataset for FDI model training.

Equations (2), (3) & (4) are used as physical constraints in constructing the loss function in the training process for the physics-informed neural network (PINN). Equation 2 is used for the filter, valve and nozzle all of which degrade via the gradual closing of DPV 1, DPV 3 and DPV 4 respectively. Since the gradual opening in DPV 5 represents a leakage in the main line, Eq. 3 is used to determine the extent of valve opening which represents the level of severity in the pipe. Equation (4) which represents the drop in gear pump volumetric efficiency when DPV 2 is opened is used in developing the loss function for the PINN model for the pump. The PINN model architecture used consists of a fully connected feedforward neural network with a Leaky version of the rectified linear unit (LeakyReLU) activation function to prevent any potential dying ReLU problem during the training process. The network has 1 input and output node, 3 hidden layers with 100 neurons in each layer. The Nadam optimizer is used for its good coverage and faster training time (Bera & Shrivastava, 2020). A Mean Squared Error (MSE) Loss function of the PINN model $L(\theta)$ used is shown below where λ is a hyperparameter manually set to 1. Figure 6 below shows the PINN model architecture for each component on the testbed. The total loss for the PINN model which consists of the data and physics loss is shown in Equation (5). Table 3 also shows the various parameters used for the PINN model and the associated loss functions in Eqs. (6), (7) & (8) where $\beta = \frac{flow \times \sqrt{SG}}{c_v}$ and N_m are treated as trainable parameters in the training process.

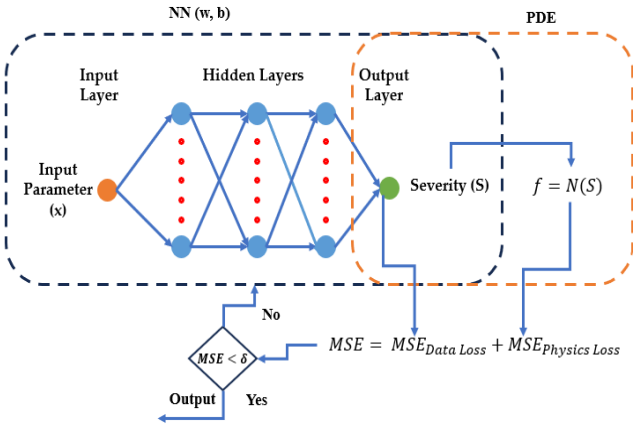


Figure 6. PINN Model Architecture for each component

$$Total\ Loss = (\lambda \times Data\ Loss) + Physics\ Loss \quad (5)$$

$$L_1(\theta) = \frac{\lambda}{N_b}\sum_{i=1}^{N_b}(S_{PINN}(x_i,\theta) - S_{obs}(x_i))^2 + \frac{1}{N_p}\sum_{j=1}^{N_p}\left(\left[V_0 - 1 + \left[\beta/\sqrt{\Delta P}\right]\right]S_{PINN}(x_j,\theta)\right)^2 \quad (6)$$

$$L_2(\theta) = \frac{\lambda}{N_b}\sum_{i=1}^{N_b}(S_{PINN}(x_i,\theta) - S_{obs}(x_i))^2 + \frac{1}{N_p}\sum_{j=1}^{N_p}\left([V_0 - 1 + N_m.N_v]S_{PINN}(x_j,\theta)\right)^2 \quad (7)$$

$$L_3(\theta) = \frac{\lambda}{N_b}\sum_{i=1}^{N_b}(S_{PINN}(x_i,\theta) - S_{obs}(x_i))^2 + \frac{1}{N_p}\sum_{j=1}^{N_p}\left(\left[V_0 - \left[\beta/\sqrt{\Delta P}\right]\right]S_{PINN}(x_j,\theta)\right)^2 \quad (8)$$

Table 3. Parameters used for the construction of the PINN Model

Component	Learning Rate	λ	Input	Output	$L(\theta)$
Filter	1e-3	1	$\Delta P = \lvert P_2 - P_1 \rvert$	$S(0,1)$	$L_1(\theta)$
Valve	1e-3	1	$\Delta P = \lvert P_3 - P_4 \rvert$	$S(0,1)$	$L_1(\theta)$
Nozzle	1e-3	1	$\Delta P = \lvert P_5 - P_4 \rvert$	$S(0,1)$	$L_1(\theta)$
Pump	1e-3	1	N_v	$S(0,1)$	$L_2(\theta)$
Pipe	1e-3	1	$\Delta P = \lvert P_5 - P_4 \rvert$	$S(0,1)$	$L_3(\theta)$

3.3.2. Approximation Model

A heuristic model $S(0,1) = 1 - x_{O.C}$ is used to approximate the level of severity of both blockages and leakages (see Eqs. 10 & 11) in the system with a domain of [0,1]. The variable x is the feature of the component which is sensitive to a change in degradation levels and it is defined as $x_{O.C}$ (see Eq. 9) with a domain of $x_{O.C}[0,1]$. The operating condition in this case is the speed of the pump.

$$x_{O.C} = \frac{Feature_{Operating\ Condition}}{Feature_{Healthy\ Condition}} \quad (9)$$

$$x_{O.C}(Blockage) = \frac{\left(Downstream\ Pressure/Upstream\ Pressure\right)_{Operating\ Condition}}{\left(Downstream\ Pressure/Upstream\ Pressure\right)_{Healthy\ Condition}} \quad (10)$$

$$x_{O.C}(Leakage) = \frac{\lvert Downstream\ Pressure - Upstream\ Pressure\rvert_{Operating\ Condition}}{\lvert Downstream\ Pressure - Upstream\ Pressure\rvert_{Healthy\ Condition}} \quad (11)$$

3.3.3. Recurrent Neural Network (RNN) Model

An FDI classifier based on a neural network architecture in a previous publication Barimah et. al (2023) which uses a recurrent neural network (RNN) architecture is used in this report. The RNN model comprises a single hidden layer with 150 neurons followed by a dense layer and a sigmoid activation function (see Figure 7 below). The model is compiled with binary cross-entropy loss and the Nadam optimizer. Early stopping is then employed to prevent overfitting during the training of the model.

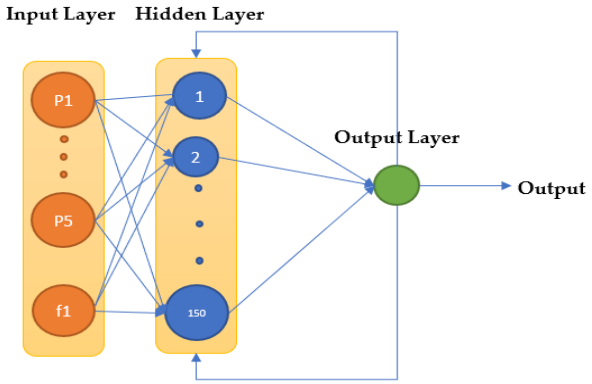


Figure 7. Recurrent Neural Network Architecture (Barimah et. al 2023)

3.3.4. PINN enabled Hybrid FDI Model

The physics-informed Neural Network (PINN) enabled hybrid FDI algorithm shown in Figure 8 is a weighted ensemble of the outputs of the RNN model, approximation model and PINN model. The weights of the model are skewed more towards the PINN model due to the limitations of purely data-driven model in the face of limited training data and its ability to generalize outside its training distribution. This PINN enabled hybrid model is then benchmarked against the other FDI algorithms, presented in Figure 4, for the system undergoing multi-component degradation scenarios. The model weights (W_D, W_P, W_A) for each component in the hybrid ensemble model are shown in Appendix B.

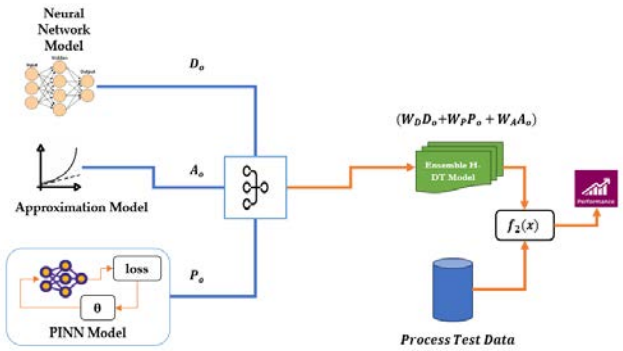


Figure 8. Ensemble Hybrid Framework for FDI models

3.3.5. Test Degradation Scenarios

Benchmarking is done using a series of test datasets recorded by Barimah et. al (2023) to assess the performance of FDI algorithms in the context of multi-component degradation. Table 3 below describes the nature of the test data with the components under consideration for the multi-component degradation scenarios, at different pump speeds, as well as their specific levels of severity.

Table 3. Test Degradation scenarios for FDI Model Testing

Dataset	Degradation (Component 1)	Degradation (Component 2)	Operational Speed Range (RPM)
T13	Pump (Medium-severity) at 45% DPV2 opening	Constant degradation of Nozzle: 30-70% DPV1 opening	700
T14	Pump (Medium-severity) at 45% DPV2 opening	Constant degradation of Nozzle: 30-70% DPV4 opening	950
T15	Filter (High severity) at 32% DPV1 opening	N/A	700 to 950
T16	Pump (Medium-severity) at 50% DPV2 opening	Nozzle (Medium-severity) at 40% DPV4 opening	700 to 950
T17	Constant degradation of Pump: 0-100% DPV2 opening	Constant degradation of Pipe: 0-100% DPV5 opening	800
T18	Intermittent faults for the pump between 45%-60% DPV2 opening	N/A	850
T19	Constant degradation of Pump: 0-100% DPV2 opening	Constant degradation of Valve: 30-70% DPV3 opening	850
T20	Pump (Medium-severity) at 55% DPV2 opening	Nozzle (High severity) at 30% DPV4 opening	700 to 950

4. RESULTS

4.1. Healthy Condition (HC) Scenario

The FDI algorithms showed very good performance in determining the healthy condition scenario in a situation where no fault had been injected into the system. Figure 9 shows the performance of all the FDI models in a healthy condition scenario with the pump speed at 700rpm and at 950 (see Appendix B). However, the performance of some of the models for components in a healthy state deteriorates once failure is introduced into the system. The performance of the FDI algorithms in faulty condition scenarios are presented in sections 4.2 to 4.4 showing the prediction of the conditions of various components on the testbed for the test scenarios.

4.2. Statistical Process Control (SPC)

The statistical process control (SPC) which relies on deviation from the mean of a process variable generally showed poor performance in the detection of the test MCD scenarios for components where faults were injected. For the test degradation scenario T13 which is has a leakage in the pump at DPV 2 of 45% opening and the gradual closure of DPV 4 which represents nozzle from 70% to 30%, the SPC model resulted in a 0.42 and 0.61 model performance for the pump and nozzle respectively (see Figure 9). In the case of T14 which has the same components under consideration but at a higher pump speed of 950rpm, the SPC showed an even poorer performance than in the case of T13 with 0.22 and 0.55 for the pump and nozzle respectively. However, for the components which had no failure injection, the SPC had a performance of 1.0 for the components not undergoing any form of degradation. This pattern of poor performance for components undergoing MCD scenarios and healthy components is seen in the rest of the test degradation scenarios T15, T16, T17, T18, T19 & T20 (see Appendix A).

4.3. Ensemble (Classifiers) and Recurrent Neural Network (RNN)

The ensemble classifier which uses the weighted outputs from logistic regression, support vector machine and decision tree classifier models also showed poor performance particularly for components not undergoing any form of degradation. This was revealed in T13 where it had a prediction performance of 0.42 for the pipe even though the pipe had no leak. This is also identified in T14 where the

model performance was 0.23 and 0.22 for the filter and pipe respectively. For components undergoing MCD scenarios the model showed some good performance for components (see Figure 10 &11). The recurrent neural network (RNN) also showed a similar pattern of prediction to the SPC model albeit slightly better than the former. The performance of the RNN model of the nozzle deteriorates from 0.81 for T13 (700rpm) to 0.45 for T14 (950rpm). This drop in performance is also seen in the pump where the performance reduces from 0.42 to 0.22. For the test degradation scenarios T15, T16, T17, T18, T19 & T20 (see Appendix), the RNN model showed very good prediction for the components not undergoing degradation. Nonetheless, for the components undergoing the MCD scenarios, the RNN model showed mixed model prediction performance.

4.4. PINN enabled Hybrid FDI Model

The performance of the PINN enabled hybrid FDI model on the test degradation scenarios in Table 3 above showed improved performance compared to the other algorithms in the context of the MCD scenarios. Although the PINN model performs better than the other FDI models in the hybrid model, it sometimes underperforms as seen in T17 (see Appendix A4) where the weighted ensemble hybrid model compensates for the limitations in the PINN model in predicting the degradation of the leak in the pipe due to the impact of the other models in the hybrid model. For all the test degradation scenarios, the hybrid approach showed a much better performance as seen in Figures 10 & 11 as well as for test scenarios T15, T16, T17, T18, T19 & T20 (see Appendix A).

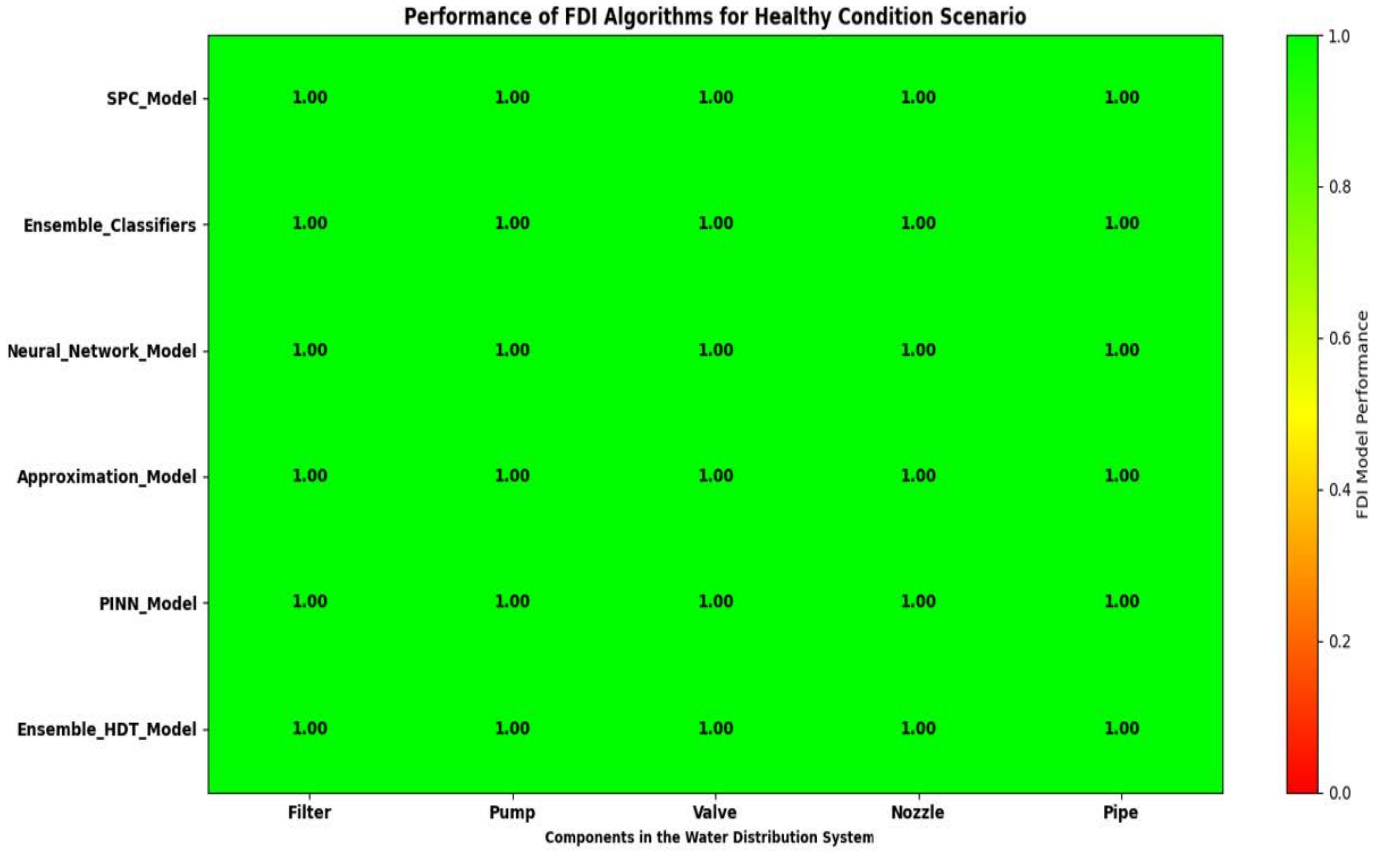


Figure 9. Performance of FDI algorithms for a Healthy Condition scenario at a pump speed of 700rpm.

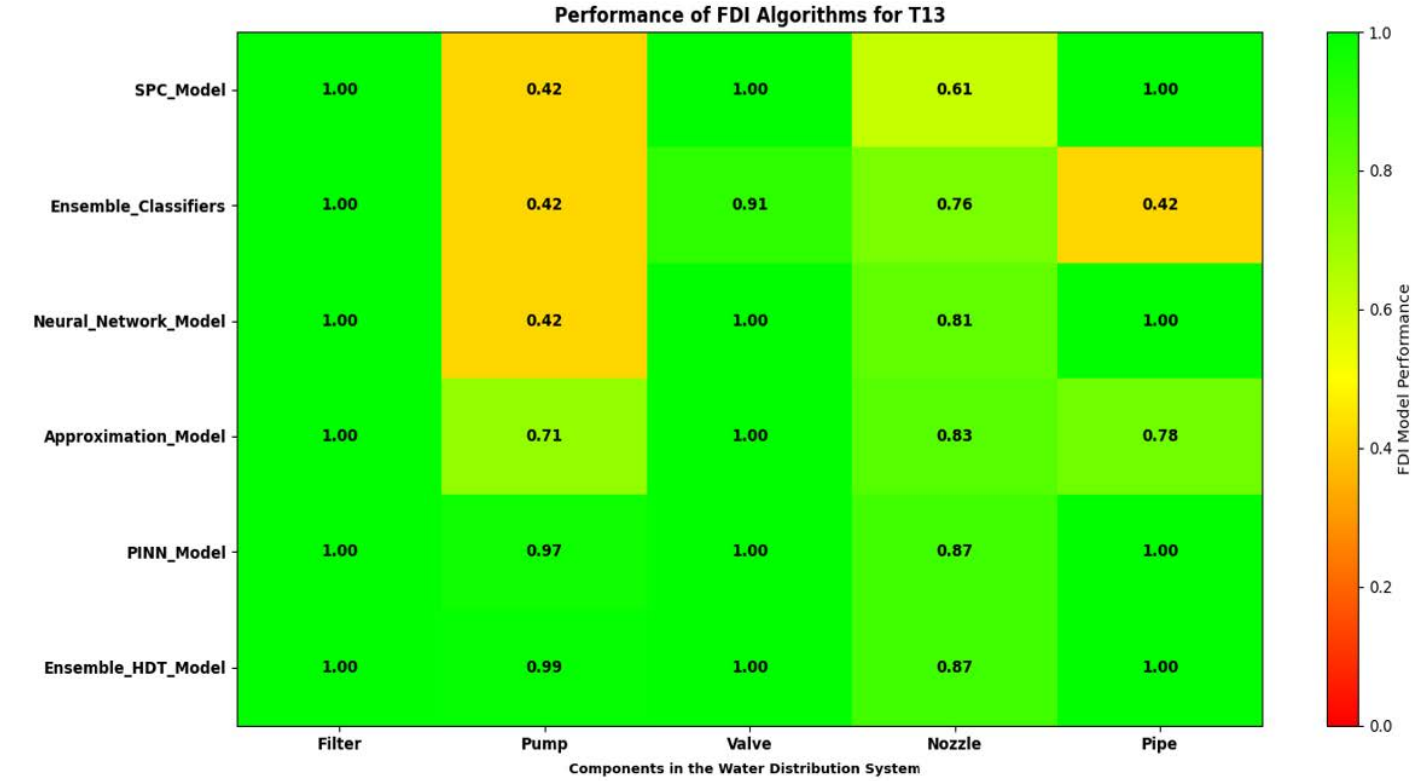


Figure 10. Performance of FDI algorithms on Test Degradation Scenario T13

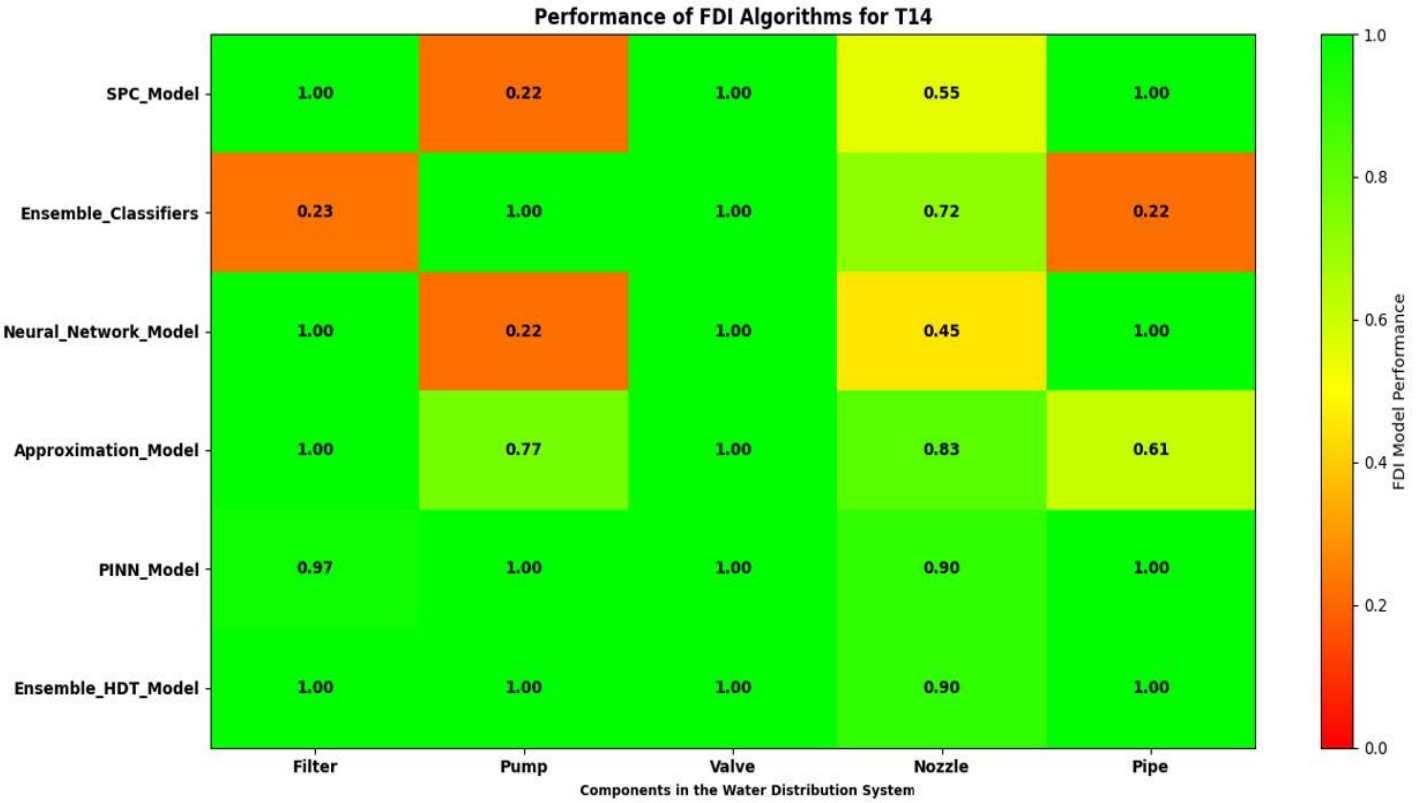


Figure 11. Performance of FDI algorithms on Test Degradation Scenario T14

5. DISCUSSION

For stakeholders in industry, training Fault Detection and Isolation models for PHM applications is key to optimizing asset health and logistics management. The challenge in model development as alluded to above is the availability of degradation data. The limited data available for training the FDI models presents a challenge in detecting MCD scenarios as seen in the results section above. The performance of the deep learning model for the test degradation scenarios showed the limitation in developing data-driven models on limited training data. This is seen in prediction of the state of the pump and nozzle in T13 & T17 as compared to the approximation and PINN models. However, it performs well when the training distribution of the available limited degradation data fall within the test degradation data scenario. This presents a challenge for stakeholders who have training data outside the distribution of the real time data from their assets. The ensemble hybrid approach proposed in Figure 8 compensates for this shortfall by integrating a heuristic approximation and a PINN approach with a neural network model to improve the overall model diagnostic performance. The main contributing parameters to the ensemble performance are the weights which were assigned using domain knowledge on the performance of the individual models with limited degradation data. This presents an interesting research opportunity for dynamically optimizing the weights in the ensemble hybrid model. The hybrid model also reduces the computational requirements for training the FDI models which ultimately reduces the cost for FDI model development for PHM applications.

6. CONCLUSIONS AND FUTURE WORK

In conclusion, this study highlights the capabilities of physics enabled fault detection and isolation algorithms for PHM diagnostics, emphasizing the challenges associated with limited training data and generalization issues. The proposed PINN-enabled hybrid model demonstrates promising FDI predictive capability for MCD diagnostics despite limited training data, indicating its potential for addressing the identification of multiple degraded conditions occurring simultaneously in a complex system. The contributions of the paper are:

C1. This study contributes to the application of physics informed FDI models for PHM applications in MCD scenarios, ultimately reducing model training data requirements for asset health management.

C2. The paper also presents an ensemble FDI approach with the capability of addressing the limitations of integrating both data-driven and physics based FDI models in multi-component degradation scenarios which can also be used in the analytics that drive digital twin applications.

Future research would focus on dynamically optimizing ensemble hybrid model weights, leveraging prediction uncertainty to further enhance model performance.

NOMENCLATURE

DT	Digital Twins
DPV	Direct Proportional Valve
FDI	Fault Detection and Isolation
$L(\theta)$	Loss Function
MCD	Multi- Component Degradation
NN	Neural Network
PHM	Prognostic and Health Management
PINN	Physics Informed Neural Network
RNN	Recurrent Neural Network
SCD	Single Component Degradation
SPC	Statistical Process Control

REFERENCES

Bararnia, H., & Esmaeilpour, H. (2022). On the application of physics informed neural networks (PINN) to solve boundary layer thermal-fluid problems. *International Communications in Heat and Mass Transfer.*

Barimah, A., Niculita, I.-O., McGlinchey, D., & Cowell, A. (2023). Data-quality assessment for digital twins targeting multi-component degradation in industrial internet of things (IIoT)-enabled smart infrastructure systems. *Applied Science, 13*(24).

Barimah, A., Niculita, O., McGlinchey, D., & Alkali., B. (2021). Optimal Service Points (OSP) for PHM enabled condition based maintenance for oil and gas applications. 6th European Conference of the Prognostics and Health Management Society.

Bera, S. and Shrivastava, V.K., 2020. Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification. International Journal of Remote Sensing, 41(7), pp.2664-2683.

Cai, S., Mao, Z., Wang, Z., Yin, M., & Karniadakis, G. (2021). Physics-informed neural networks (PINNs) for fluid mechanics. *A review. Acta Mechanica Sinica*, 1727-1738.

Duriez, T., Brunton, S., & Noack, B. (2017). *Machine learning control-taming nonlinear dynamics and turbulence.* Cham: Springer.

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., & Ryne, R. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing, 26*(2), 448-466.

Hu, Y., Miao, X., Si, Y., Pan, E., & Zio, E. (2022). Prognostics and health management: A review from the perspectives of design, development and decision. *Reliability Engineering & System Safety, 217*, 108063.

Huang, B., & Wang, J. (2022). Applications of physics-informed neural networks in power systems-a review. *IEEE Transactions on Power Systems, 38*(1), 572-588.

Knight, E., Russell, M., Sawalka, D. and Yendell, S., 2013. ValveModeling. Control Valve Wiki.

Lu, Q., Xie, X., Parlikad, A., & Schooling, J. (2020). Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. *Automation in Construction, 118*, 103277.

Maass, W., Parsons, J., Purao, S., Storey, V., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *ournal of the Association for Information Systems,, 19*(2), 1.

Rizi, S., & Abbas, M. (2023). From data to insight, enhancing structural health monitoring using physics-informed machine learning and advanced data collection methods. *Engineering Research Express, 5*(3), 32003.

BIOGRAPHIES

Atuahene Barimah is an Operations Assistant at Wellscope Energy Solutions and is currently a PhD researcher at Glasgow Caledonian University where he received his MSc. in Applied Instrumentation and Control at in 2020. He also received his BSc in Petroleum Engineering at the Kwame Nkrumah University of Science and Technology in 2017. His research interests include data-driven maintenance, system reliability, process control, project management, operations research, digital twin design, IIoT, and computational finance.

Octavian Niculita is a Senior Lecturer in Instrumentation with Glasgow Caledonian University. He has a PhD in Industrial Engineering from the Technical University of Iasi, Romania carried out under the EDSVS framework. His current research interests include industrial digitalization, predictive maintenance, PHM system design, integration of PHM and asset design for aerospace, maritime, and oil & gas (surface and subsea) applications. Octav has over ten years of experience in design and development of prognostics and health management applications, having worked on applied aerospace projects funded by The Boeing Company and BAE Systems as a Research Fellow and Technical Lead on his previous appointment with the IVHM Centre at Cranfield University, UK. He is a member of the Prognostics and Health Management Society, InstMC and the IET.

Don McGlinchey graduated from Strathclyde University with a BSc (Hons) Physics before working as a project engineer at Babcock Energy Ltd. He returned to academia and gained an MSc in Bulk Solids Handling Technology and his Doctorate on a study of the effect of vibration on powder beds. He is currently a Professor in the Department of Engineering at Glasgow Caledonian University where he is the academic leader in teaching, research, and consultancy in the area of multi-phase flow. He has edited two books, and authored over 100 papers, articles, and consultancy reports.

Andrew Cowell is the Chair of the Department of Engineering Industrial Advisory Group at Glasgow Caledonian University. In addition to his short-term contracts in research into coal handling for entrained flow gasifiers, and particulate solids handling education materials, Andrew has undertaken consultancy projects in the United Kingdom (UK), and delivered short courses in the USA, Sweden and the UK. He has also presented at international academic conferences in the UK, Australia, Norway and Spain. He is a Chartered Engineer, a Member of the IMechE and Fellow of the Institution of Engineers and Shipbuilders in Scotland.

Billy Milligan is the Director of Instrumentation and Control at Chart Industries. He has worked primarily on the design, testing and commissioning of screw compressor control systems within the oil and gas industry for the past 15 years. He is a member of the InstMC and the IET.

APPENDIX

Appendix A. FDI Model Performance

Figure A1. Healthy condition at 950 RPM

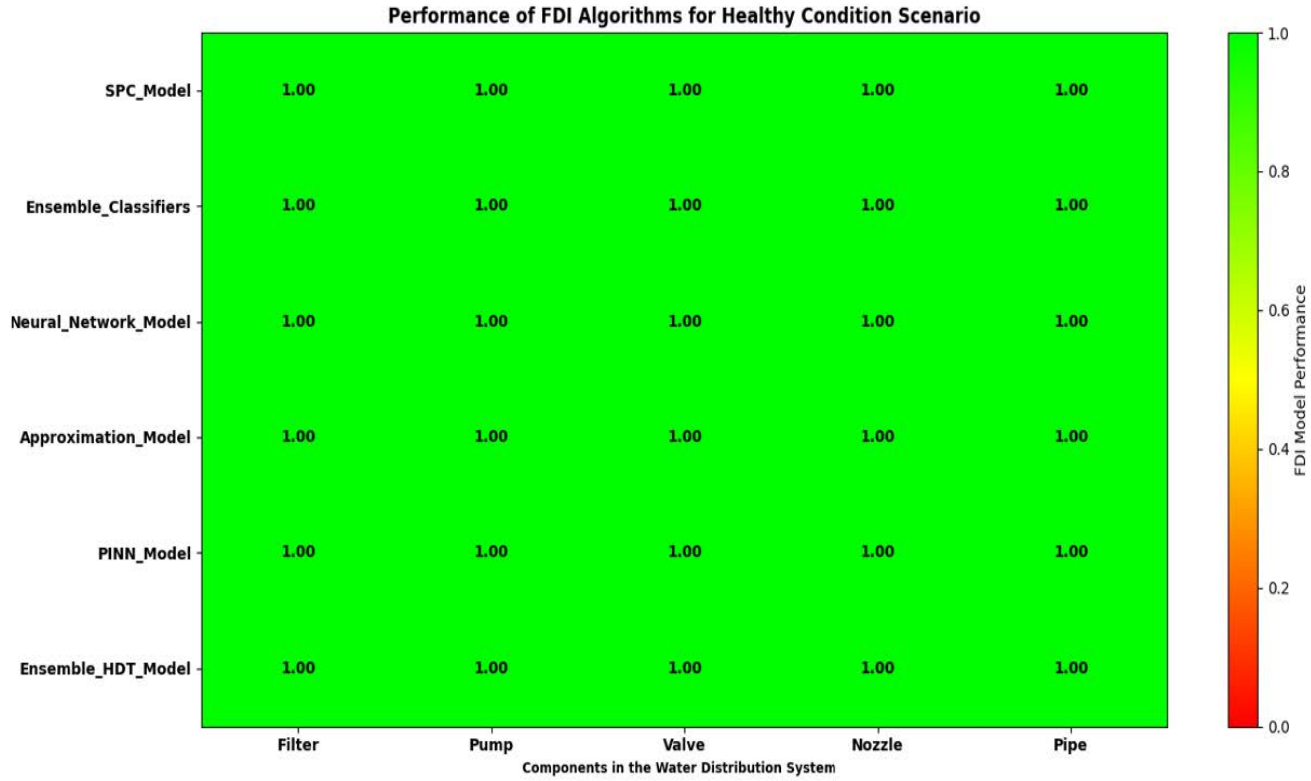


Figure A2. Test Scenario T15

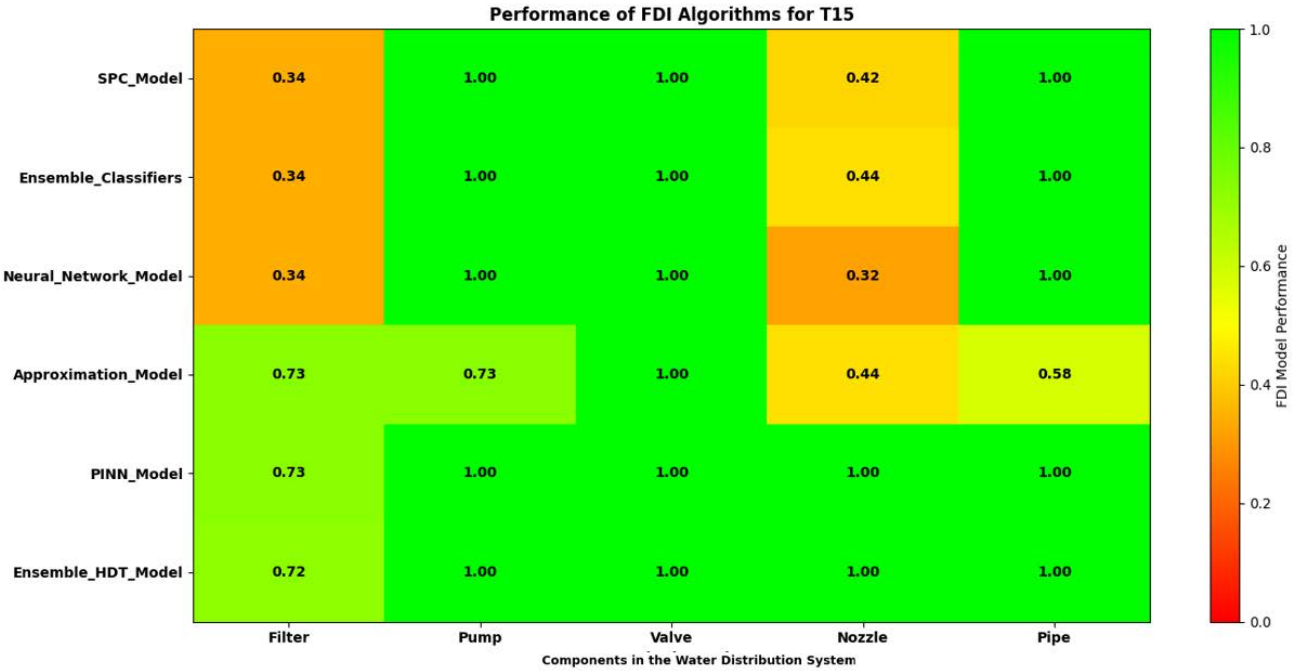


Figure A3. Test Scenario T16

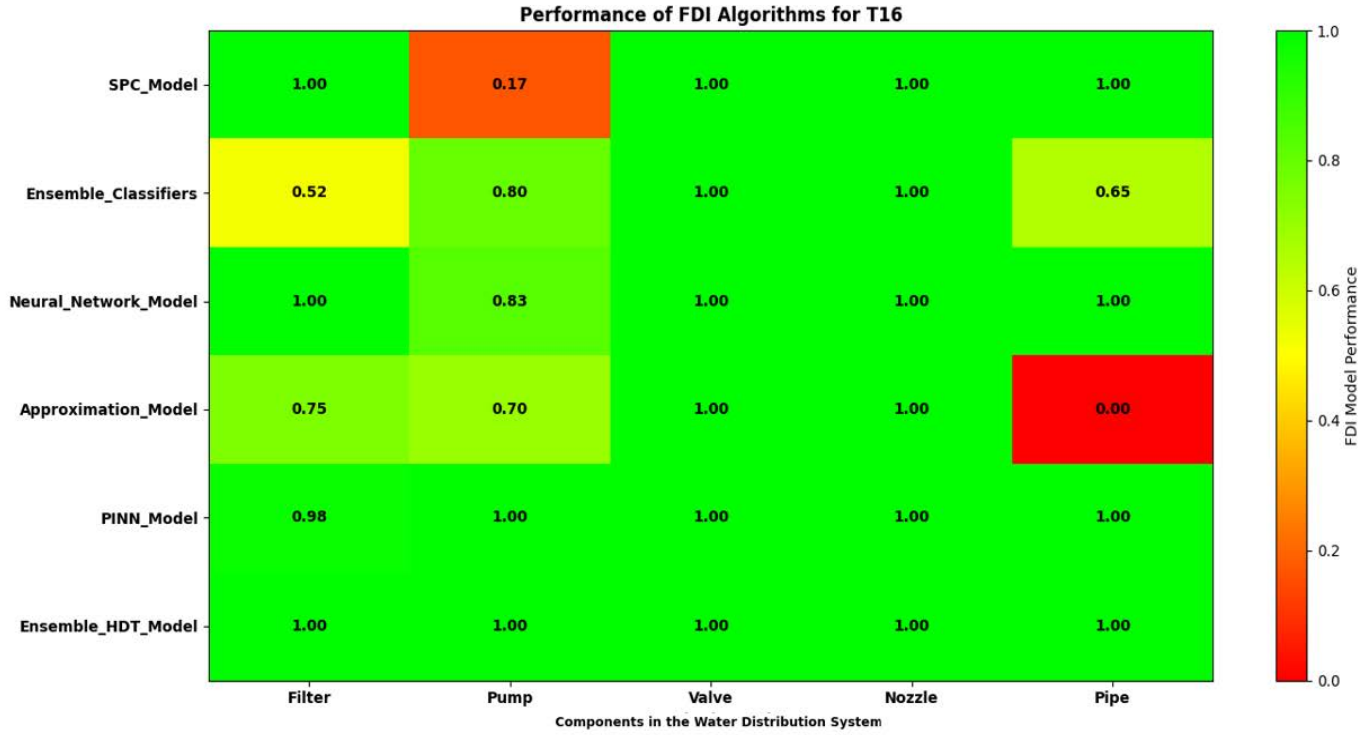


Figure A4. Test Scenario T17

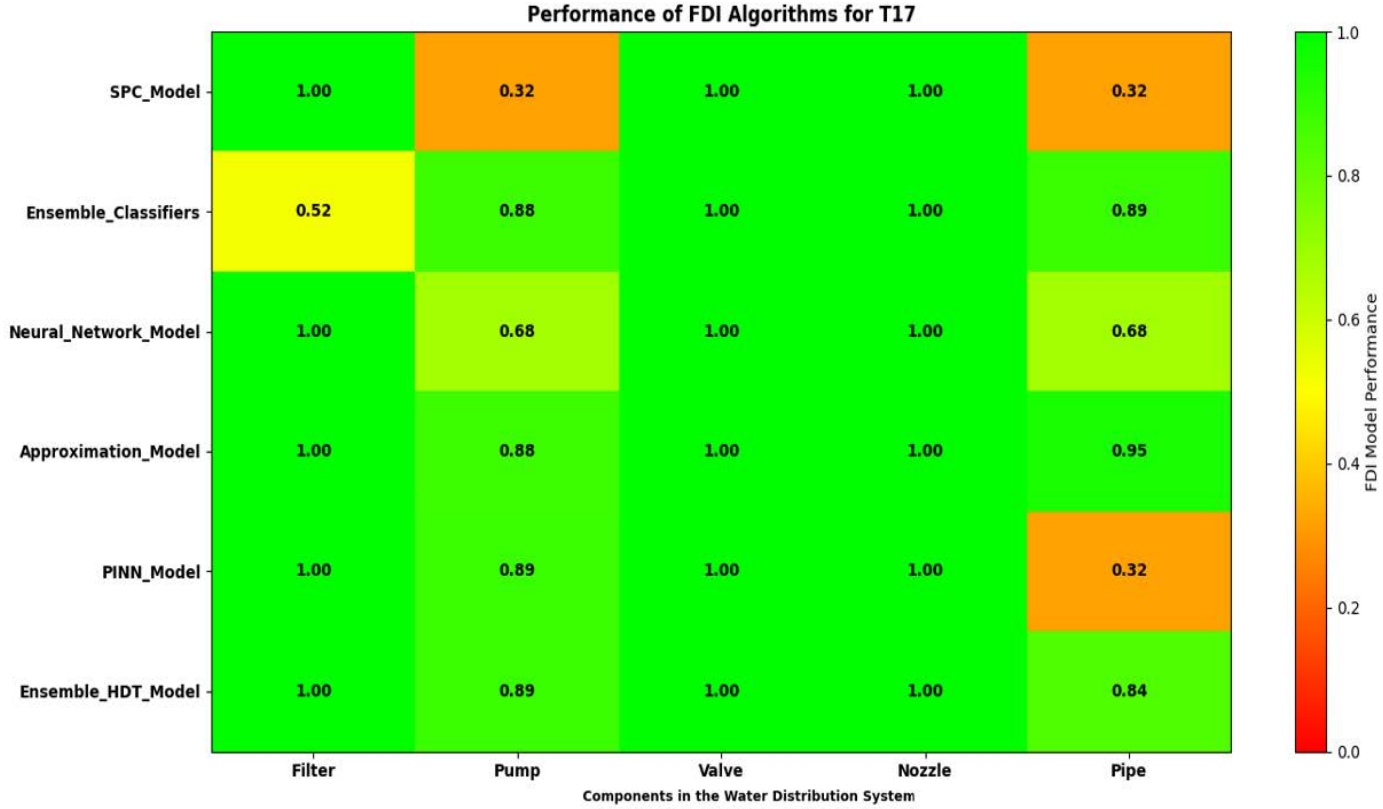


Figure A5. Test Scenario T18

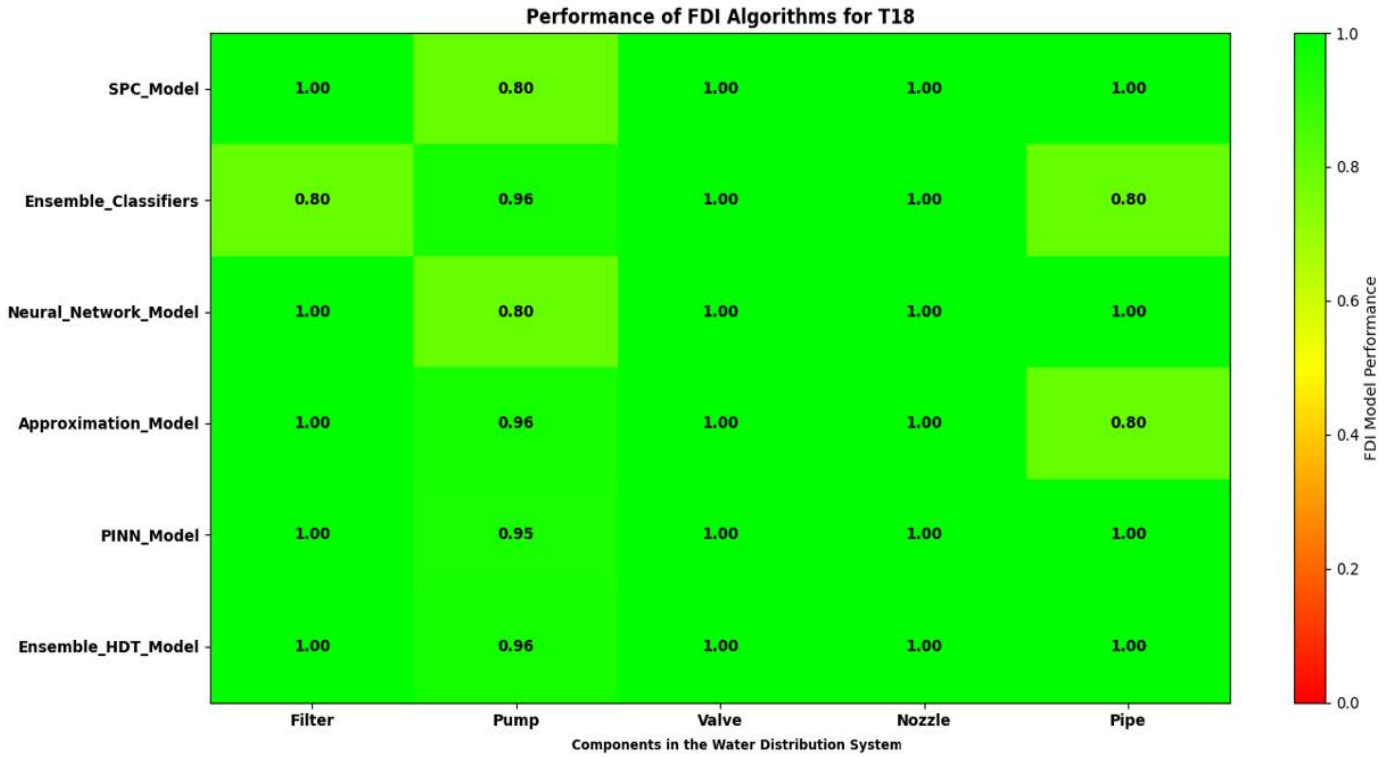


Figure A6. Test Scenario T19

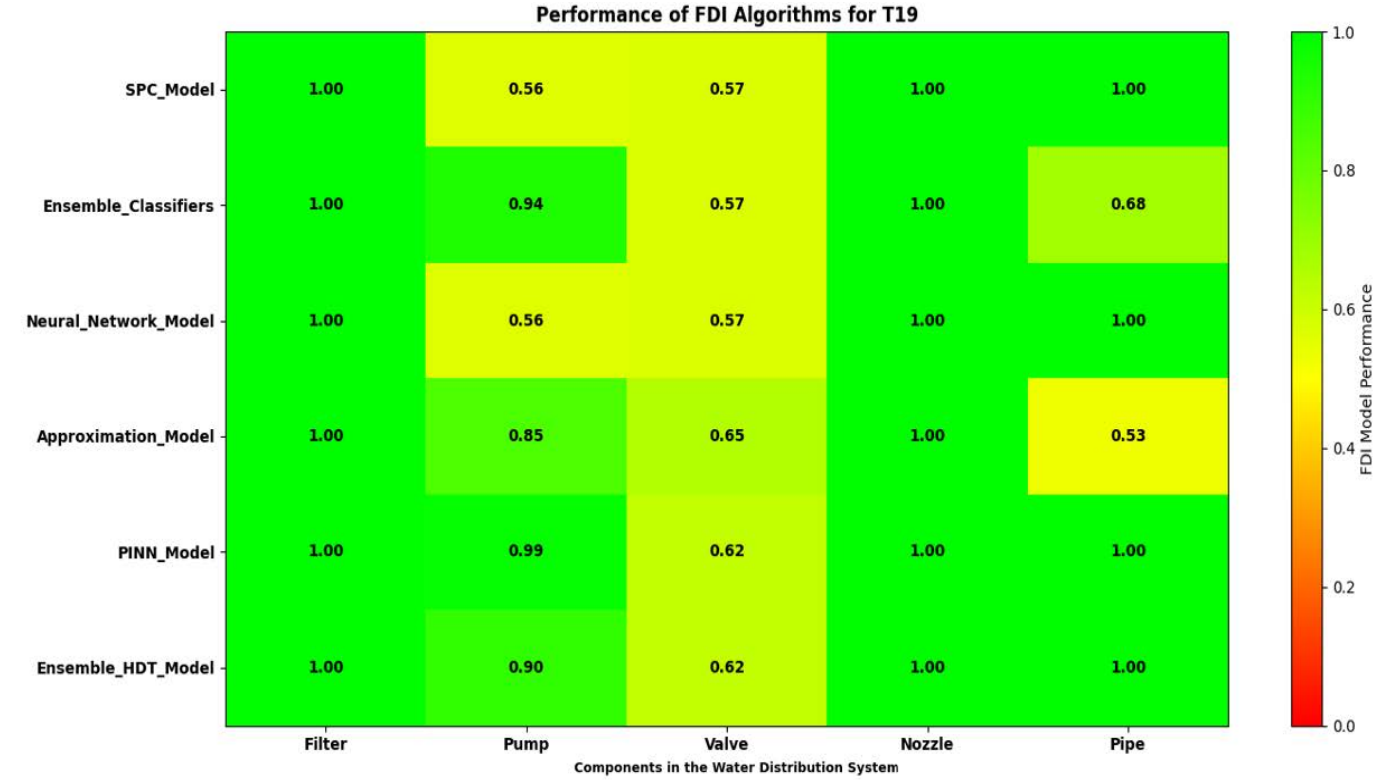
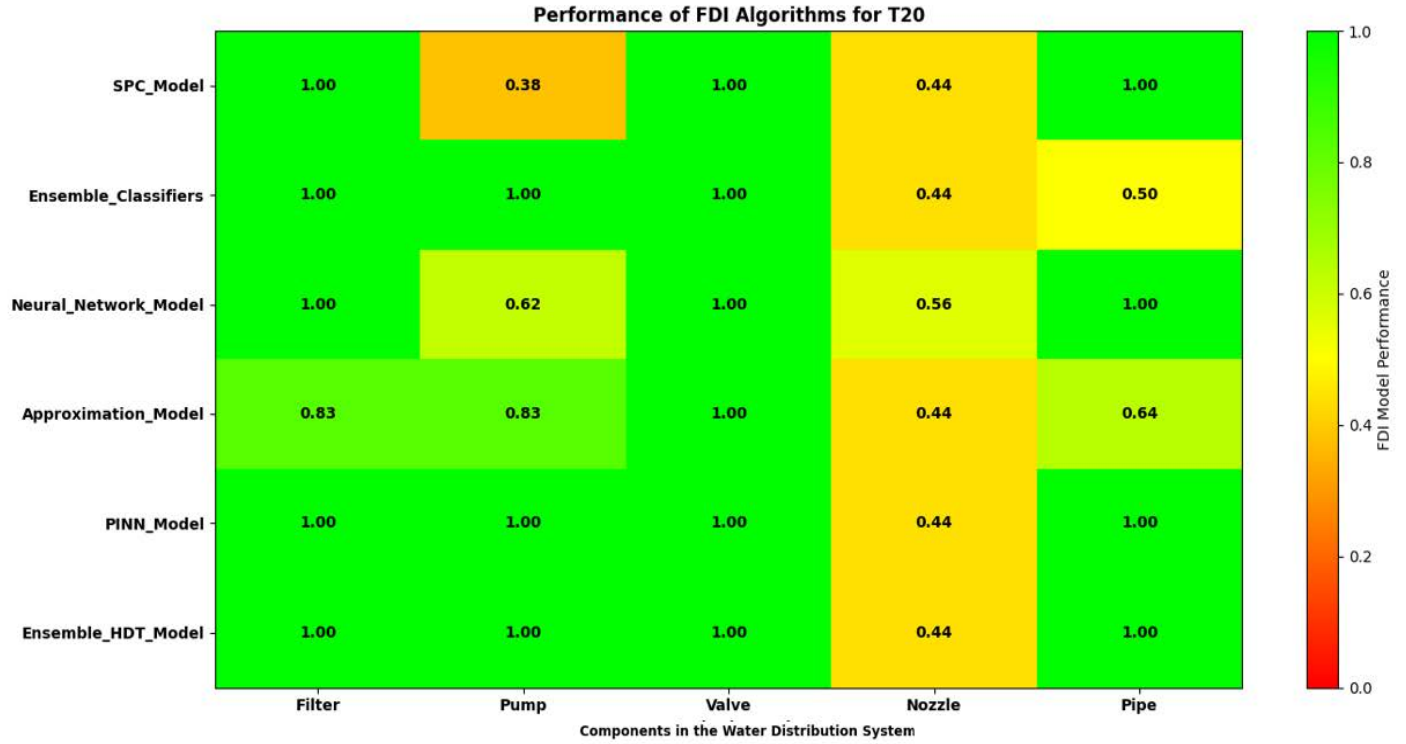
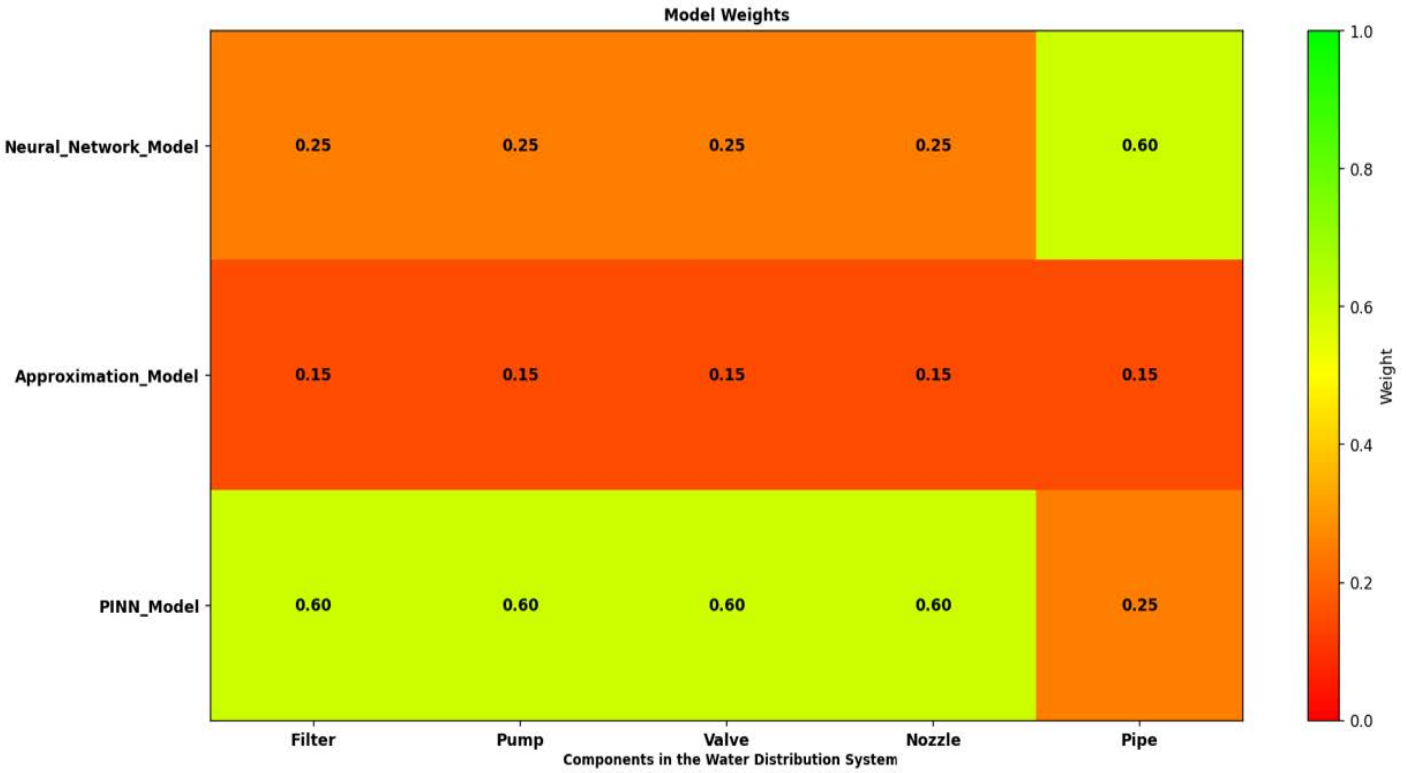


Figure A7. Test Scenario T20



Appendix B. Model Weights for FDI algorithms



Transfer Learning-based Adaptive Diagnosis for Power Plants under Varying Operating Conditions

Jiwoon Han[1], and Daeil Kwon[2]

[1,2]*Department of Industrial Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do, 16419, South Korea*
jiwoon98@g.skku.edu
dikwon@skku.edu

ABSTRACT

Transfer learning is a method that transfers knowledge learned from a source domain to a similar target domain to improve learning. In power plants, obtaining sufficient anomaly data is difficult due to the characteristics of the systems. Transfer learning enables learning with only a small amount of data from the target domain by using a model trained in a similar domain. By applying transfer learning, models developed for one power plant can be expanded and used in other power plants where available data are limited.

Using actual data from an operating combined-cycle power plant, an anomaly diagnosis model was developed and tested. Its applicability to different operating conditions and anomaly cases was evaluated through transfer learning. The fine-tuned pre-trained model was effectively adapted with limited target domain data. Transfer learning was applied despite the limitations of data and distribution differences. The expandability of anomaly diagnosis models to different power plant systems was demonstrated by applying transfer learning.

1. INTRODUCTION

The limited anomalous data and labels in power plants are challenges for training anomaly diagnosis models. Due to the requirements for safety and operational stability, inducing failures or obtaining sufficient anomalous data is difficult in power plants (Qian & Liu, 2023). Variations in operating conditions also complicate model training by changing the distribution of data. The operating conditions of power plants change with variations in power demand over time and external factors such as temperature and humidity (Bai, Yang, Liu, Liu, & Yu, 2021). In actual operating power plants, it is difficult to obtain data while operating under the same conditions consistently, as power demands and external factors vary. Differences in operating conditions disrupt the assumption of consistent data distribution between training and testing sets in anomaly diagnosis models (Li, Lin, Li, & Wang, 2022; Zhou, Lei, Zio, Wen, Liu, Su, & Chen, 2023).

Developing diagnosis models for a new power plant system incurs additional costs, even after significant investments have been made to overcome challenges and develop the models. This is because the distribution of data collected varies due to differences in the structure and sensors of the systems in each new power plant. Each new power plant requires a customized approach to model development, involving the redesign of diagnosis models to fit the specific data characteristics of that plant. To develop models for other new power plants, the process should start anew with data collection. Training and validating models with the collected data are essential steps in developing the new model. This process again incurs significant time and costs.

The fact that power plants of the same type share a common domain can be utilized. When applying models to new power plants, it is typically necessary to redesign them due to differences in data distribution. Since the power plants operate on similar principles within the common domain, this can enable the expansion of existing models without a complete redesign. This approach utilizes the commonalities from the same types of plants, reducing development time and costs.

By applying transfer learning, a developed model can be expanded and adaptively used for a new power plant within a similar domain. Transfer learning is a method that transfers knowledge learned from a source domain to a target domain with insufficient data for a similar task (Pan & Yang, 2009). The transfer learning method involves fine-tuning model parameters pre-learned from the source domain using limited data from the target domain. With transfer learning, a model developed in the source domain can be adapted to a new system in the target domain, instead of restarting the entire process. Additionally, it can be applied to the target domain using only a small amount of data, serving as an approach to overcome the challenges of limited data and labels. By

applying transfer learning and using the model adaptively, the expandability and practicality of the diagnosis model can be enhanced.

In this paper, an anomaly diagnosis model was developed using data collected from the actual gas turbine of an operating combined-cycle power plant. The developed diagnosis model was tested by applying transfer learning to data with different anomaly features and operating conditions than the training data. Collected data have an emergency shutdown called a "trip", that occurs in the case of anomalies to prevent serious accidents. Cases, where actual data are collected under different operating conditions, have similar situations with other power plants data that have different data distributions. By fine-tuning with limited data from the target domain, this study demonstrated the potential to expand a developed model to different power plant systems. Comparative analysis was conducted by applying transfer learning, even in situations of data imbalance where little anomaly data is available in the target domain.

Section 2 introduces the related works that developed a diagnosis model for the power plant and applied transfer learning to the model. Section 3 describes the data, model, and transfer learning methods used in this study. Section 4 presents the results, Section 5 discusses these results, and Section 6 presents the conclusions and future work.

2. RELATED WORKS

Related studies on anomaly diagnosis in power plants have been conducted across various subjects and domains. Diagnosis using the Gaussian Process (GP) algorithm and model ensemble techniques were conducted at an actual coal-fired thermal power plant (Zhang, Dong, Kong, & Meng, 2019). They identified relationships between variables to reflect temporal dependencies and cross-variable associations, using combinational data relationships to develop the diagnosis model. Lee et al. (2021) collected data from a full-scope simulator for abnormality diagnosis in a nuclear power plant and developed a Convolutional Neural Network (CNN) algorithm model. To manage the 1004 sensor variable data, they converted it into two-channel 2D images with a data size of 32*32.

As mentioned in the introduction, power plants have challenges due to the limited anomaly data and differences in operating conditions. To address these challenges, transfer learning methods have actively been researched for diagnosing power plants. Studies have been conducted to apply transfer learning for fault diagnosis at different power levels in nuclear power plants. Data were collected at several power levels using a simulator, and a CNN algorithm was developed to handle numerous sensor variables. Maximum Mean Discrepancy (MMD) was used to develop the model to adapt to differences in distributions when power levels vary. With these approaches, Li et al. (2022) divided domains based on power levels and applied transfer learning across different power levels. They also analyzed the effects of various kernel functions used to calculate MMD. Wang et al. (2022) utilized Transfer Local MMD (TLMMD) combined with the ResNet-18 algorithm to develop a diagnosis model. Li, Lin, Li, and Wang, (2022) applied transfer learning to construct models for each power level. They proposed a framework that determines the current power level during actual operation and matches data to the model trained at each power level.

The CNN algorithm and transfer learning were also applied for fault detection in the gas turbine combustion chambers of power plant systems (Bai et al., 2021). Exhaust Gas Temperature (EGT) data collected from two gas turbines were used. The turbine with more data was used as the source domain for training, and transfer learning was then applied to the other turbine, which had limited data. The performance of the transfer learning approach was evaluated and compared with various other diagnosis methods.

3. APPROACH

A diagnosis model was developed for the gas turbine of an operating combined-cycle power plant. Training and testing were conducted using data from collected anomaly cases, and transfer learning was applied. The model's performance was evaluated, observing changes in performance based on the data used for training and the application of transfer learning.

3.1. Data

The operating data were collected from sensors related to the gas turbine equipment of a combined-cycle power plant A, located in region B of Korea. The power plant data were provided by KEPRI (Korea Electric Power Corporation Research Institute). A total of eight anomaly cases related to trips were detected. Data for each case were collected on the dates when the anomaly occurred for four years. Each case has different operating conditions, resulting in different characteristics.

Table 1. Collection of Data.

Collection period	4 years	
Number of sensors	118	
Number of Cases	8	
Data instance per cases	Total	256
	Normal	128
	Anomaly	128

Data were collected from 118 sensors of the plant's gas turbine system. Sensors collected data on flow rate, pressure, and temperature, such as EGT. Each sensor was related to the control and flow of fuel gas in the gas turbine.

Within each of the eight anomaly cases, there are 128 instances of both normal and anomaly data, labeled by

domain experts based on the investigation reports conducted for each case. The data format consists of 60-minute windows for the 118 collected sensor data points.

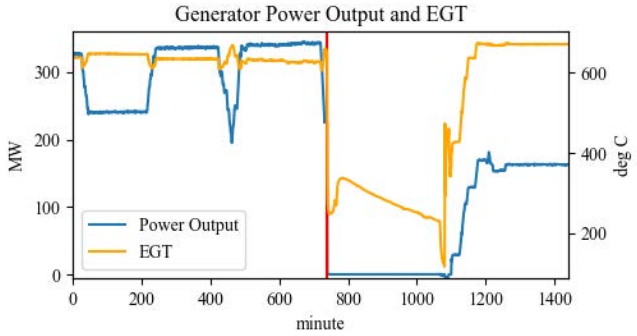


Figure 1. Example of collected data with a trip.

To utilize the overall 118 sensor data, the collected time series data can be concatenated in parallel to form a two-dimensional matrix. Each row is composed of a time series, and the patterns of the sensors contain information about anomalies. The data from 118 sensors have variations in units and ranges of values, depending on their measurement targets. To address this, min-max normalization was applied to each sensor. The matrix collected from 118 sensors over a 60-minute window is represented as an image as follows.

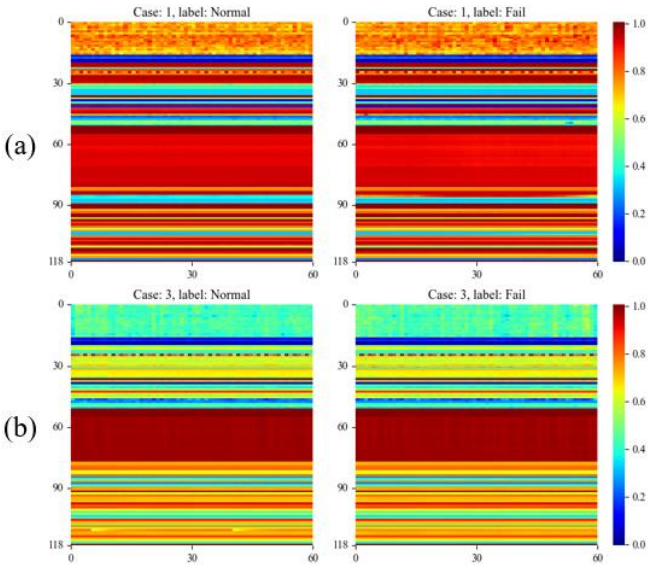


Figure 2. Example image of data. (a) Case 1, (b) Case 3.

3.2. Transfer Learning

Transfer learning was applied by taking a model trained in the source domain and fine-tuning it with a limited dataset from the target domain, with some weights of the convolutional layer fixed. The model was constructed using a Convolutional Neural Network (CNN) architecture, with 1D kernels utilized to detect patterns in the data from 118 sensors over a 60-minute window. The model's structure includes convolutional layers, max-pooling layers, batch-normalization layers, and fully connected layers. It is designed to learn features from the data and perform classification.

In the CNN model, the initial convolutional layers extract general features, while the fully connected layers extract specific features (Zhu, Peng, Chen, & Gao, 2019). This characteristic enables the use of general features validated in the source domain while adapting specific features for classification in the target domain when applying transfer learning. The model was trained using all available data in the source domain, and fine-tuning was conducted with only 60 data instances from the target domain.

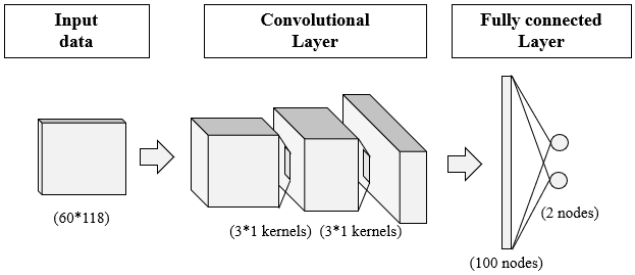


Figure 3. Outline for Structure of CNN model

3.3. Evaluation

Eight cases were used as different source domains, with each dataset being used to train a CNN model individually. Each trained source case model was evaluated using the validation data from that case and tested using the remaining seven cases as test data. Thus, there are eight evaluation results for one case model and a total of 64 results for all eight cases. The average of the calculated performance metrics was used to evaluate how each method applies. This average performance metric was compared based on the application of transfer learning and depending on the case used for training.

The performance metric used is the Matthews Correlation Coefficient (MCC), which represents performance through the correlation between actual and predicted labels, among metrics for binary classification (Chicco, Tötsch, & Jurman, 2021). MCC values range from [-1, 1], as it is a correlation coefficient. Accuracy, the commonly used performance metric, cannot represent cases of class imbalance and cases where predictions are made with only one label. The MCC metric can effectively show the relationship between actual and predicted labels. It is considered to perform well even in cases of class imbalance, indicating a value of 0 when predictions are made with only one label.

4. Result

The comparison of the average MCC for the three cases is shown in the figure below. CNN models were trained using data from eight different anomaly cases. The models for each

case were trained and validated in an 8:2 ratio, and data from different cases, which were not used in training, were utilized for testing.

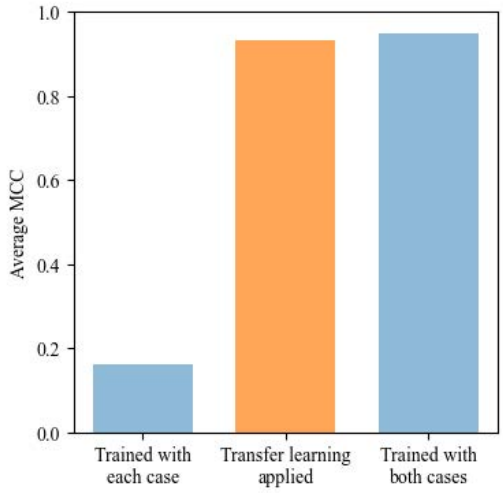


Figure 4. Comparison of the average MCC

The performance of the model significantly improved when transfer learning was applied, compared to using data from only one case. Additionally, although the best performance was in the case that used data from both domains, the performance with transfer learning exhibited similar levels of effectiveness.

Each anomaly case has different operating conditions and anomalies. While the models exhibit high performance in validation for each specific case, most exhibit low MCC scores when applied to other cases. Most test cases with low MCC scores are cases where predictions are made with only one label, either normal or anomaly. Figure 4 shows that the diagnosis models are well-trained for each specific case. Additionally, it indicates that the models cannot predict accurately in tests for other anomaly cases due to different operating conditions and anomalies.

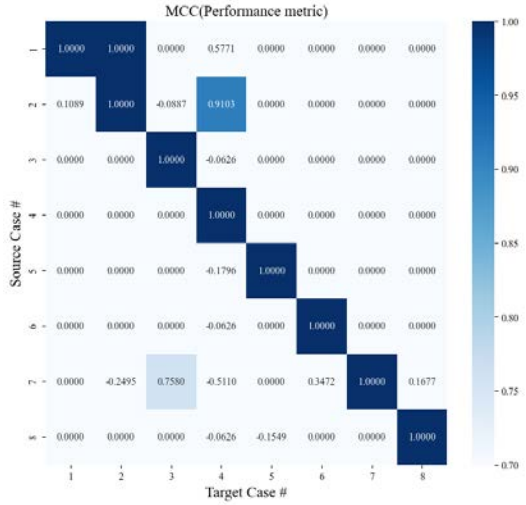


Figure 5. Performance of models trained with each case.

The performance of each case was evaluated by applying transfer learning. To evaluate the effectiveness of transfer learning, a comparison was made between models that applied transfer learning and those that used all the data from both the source and target domains without transfer learning. Initially, each case was fine-tuned with limited data from the target domain, based on the model trained in the source domain. The performance metrics were evaluated using test data that were not used in the fine-tuning of the target domain. Transfer learning was applied as described in Section 3.2.

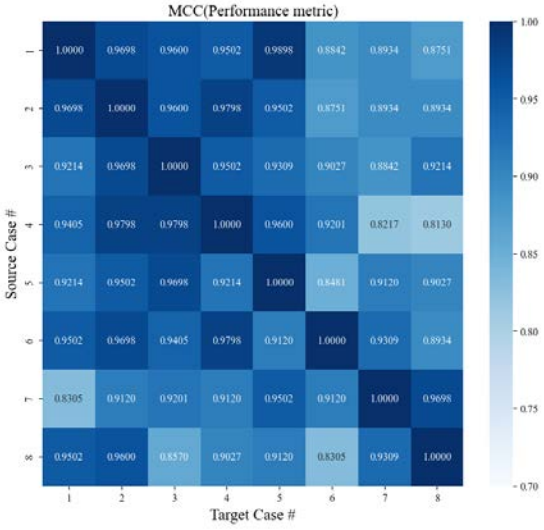


Figure 6. Performance of models applied transfer learning.

For comparison, a scenario was assumed in which data were collected and available from both domains. The model was trained using all the data from both the source and target domains without the use of transfer learning. The results are as follows.

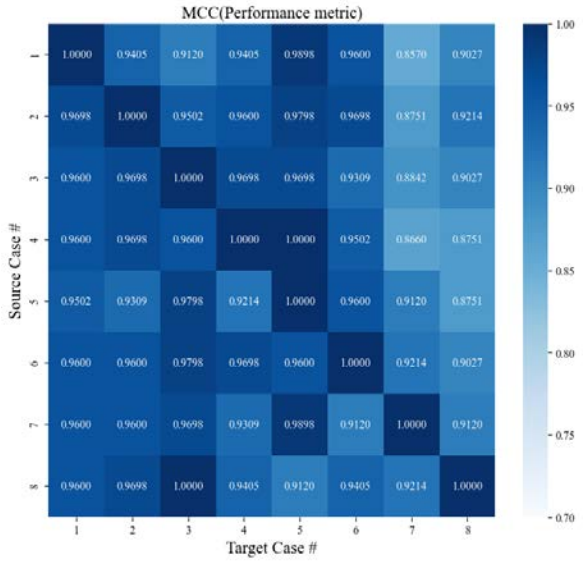


Figure 7. Performance of models trained with both domains.

When trained using data from both domains, the data for training and validation was the same as when only one domain was trained and validated, but the trained model was not completely accurate. This is because each domain has different operating conditions, making it challenging to treat and learn from them as a single domain.

It can be observed that diagnoses performed with transfer learning are effective. The difference in the effectiveness is due to the different distributions of data from each anomaly case under different operating conditions. When using models trained with a single case, the features identified during training differ from those in the test, leading to poor performance of the model. In contrast, the application of transfer learning has shown that fine-tuning the model with a limited amount of data can enhance performance. When compared to cases where data from both the source and target domains are available, similar performance metrics were observed. This indicates the effectiveness of transfer learning in cases with different data distributions. When expanding the model to different new power plant systems, data newly collected under different operating conditions or anomaly cases differ from the previously trained data. Previous results demonstrate the expandability of the diagnosis model through transfer learning, which has been effective despite these differences.

5. DISCUSSION

In real-world situations, collecting anomaly data is more challenging compared to normal data, resulting in a data imbalance. Additional analysis was conducted to monitor changes in the training models by applying transfer learning. A sensitivity analysis was performed on the ratio of normal to anomaly data used in fine-tuning the target domain. To address this, the performance of transfer learning was evaluated by gradually increasing the proportion of normal data. Figure 8 shows the results of the average MCC according to the ratio of normal to anomaly data used in fine-tuning. The results indicate a significant decrease in performance when no anomaly data were included.

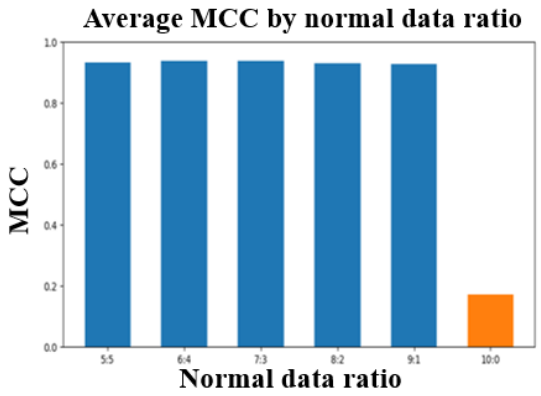


Figure 8. Average MCC according to the ratio of normal to anomaly data.

Using transfer learning, the ratio of normal to anomaly data used for fine-tuning was changed to analyze factors that influence its effectiveness. It was observed that even a small amount of anomaly data could yield good results when the overall proportion of anomaly data was reduced. In real-world application scenarios, normal data is generally more prevalent. The model could identify anomalies well even in imbalanced situations where the ratio of anomaly data was 9 to 1. However, the performance significantly decreased when there was no anomaly data at all. This decrease occurs because fine-tuning without any information about anomalies makes it difficult to adaptively use the identified anomaly features through transfer learning.

6. CONCLUSION AND FUTURE WORK

In this study, an anomaly diagnosis model was developed using data from an actual combined-cycle power plant. Diagnosis models were developed for each case, and it was observed that the operating conditions and anomaly features varied according to each case. To use the diagnosis models adaptively, transfer learning was applied to fine-tune the models and evaluate their performance. Using transfer learning, the ratio of normal to anomaly data used in the fine-tuning of the target domain was varied to analyze changes in performance. This process demonstrated that transfer learning could be effectively applied even in imbalanced situations with a predominance of normal data, and it also highlighted the importance of collecting anomaly data.

Next, Research could be conducted to validate the expandability of the model through transfer learning using data collected from different new power plants. Research could be conducted on applying the model in real-time

scenarios at actual power plants using transfer learning. In actual power plant operations, the occurrence of an anomaly is already critical. There is a need for an approach that allows for fine-tuning without information about anomalies and adaptively uses the model under different operating conditions. Additionally, considering that data are collected in a sequential time series, there is a need for a transfer learning framework that fine-tunes the model using only initial data and adaptively detects anomalies.

ACKNOWLEDGEMENT

This work was partly supported by Korea Institute of Energy Technology Evaluation and Planing(KETEP) grant funded by the Korea government(MOTIE) (20217010100020, Development of Intelligent O&M Technologies for Standard Combined Cycle Power Plant).

REFERENCES

Bai, M., Yang, X., Liu, J., Liu, J., & Yu, D. (2021). Convolutional neural network-based deep transfer learning for fault detection of gas turbine combustion chambers. Applied Energy, 302, 117509.

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData mining, 14, 1-22.

Lee, G., Lee, S. J., & Lee, C. (2021). A convolutional neural network model for abnormality diagnosis in a nuclear power plant. Applied Soft Computing, 99, 106874.

Li, J., Lin, M., Li, Y., & Wang, X. (2022). Transfer learning network for nuclear power plant fault diagnosis with unlabeled data under varying operating conditions. Energy, 254, 124358.

Li, J., Lin, M., Li, Y., & Wang, X. (2022). Transfer learning with limited labeled data for fault diagnosis in nuclear power plants. Nuclear Engineering and Design, 390, 111690.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Qian, G., & Liu, J. (2023). Fault diagnosis based on gated recurrent unit network with attention mechanism and transfer learning under few samples in nuclear power plants. Progress in Nuclear Energy, 155, 104502.

Wang, Z., Xia, H., Zhang, J., Annor-Nyarko, M., Zhu, S., Jiang, Y., & Yin, W. (2022). A deep transfer learning method for system-level fault diagnosis of nuclear power plants under different power levels. Annals of Nuclear Energy, 166, 108771.

Zhang, Y., Dong, Z. Y., Kong, W., & Meng, K. (2019). A composite anomaly detection system for data-driven power plant condition monitoring. IEEE Transactions on Industrial Informatics, 16(7), 4390-4402.

Zhou, H., Lei, Z., Zio, E., Wen, G., Liu, Z., Su, Y., & Chen, X. (2023). Conditional feature disentanglement learning for anomaly detection in machines operating under time-varying conditions. Mechanical Systems and Signal Processing, 191, 110139.

Zhu, Z., Peng, G., Chen, Y., & Gao, H. (2019). A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis. Neurocomputing, 323, 62-75.

Ultrafast laser damaging of ball bearings for the condition monitoring of a fleet of linear motors

Abdul Jabbar[1], Manuel Mazzonetto[1], Leonardo Orazi[1, 2] and Marco Cocconcelli[1]

[1] *DISMI - University of Modena and Reggio Emilia, Via Amendola 2, Reggio Emilia, 42122, Italy*

[2] *EN&TECH - University of Modena and Reggio Emilia, Piazzale Europa, 1, Reggio Emilia 42124, Italy*
abdul.jabbar@unimore.it
manuel.mazzonetto@unimore.it
leonardo.orazi@unimore.it
marco.cocconcelli@unimore.it

ABSTRACT

Machine learning-based condition monitoring of mechanical systems, such as bearings, employs two primary approaches: unsupervised and supervised methods. Unsupervised approaches aim to characterize the healthy state of the machine and monitor deviations from this state. The advantage lies in requiring only the health condition of the component without the need for historical data until breakdown. However, the disadvantage is the lack of information regarding the root cause of any potential malfunction. On the other hand, supervised methods consider both healthy and faulty cases, aiming to maximize the difference between them through post-processing, as well as among different fault types. The advantage is the ability to analyze the specific signature of a particular fault type. Nonetheless, the disadvantage is that available data usually do not cover all possible faults that may occur. Typically, obtaining a faulty bearing involves either a time-consuming run-to-failure test or the artificial induction of faults using drills, electro-discharge pens, etc. While artificial faults offer a quicker procedure, they often fail to replicate real faults faithfully. This paper suggests using picosecond laser technology to engrave the surface of the bearing and create artificial faults. Modern laser technology allows for precise control over the dimensions of injected faults, enhancing the understanding of fault progression at various stages in the life of bearings. These measurements are crucial parameters for evaluating the robustness of diagnostic algorithms. This paper focuses on artificially damaging a ball bearing used in an independent cart systems application, which comprises a fleet of linear motors moving on the same rail. These systems have recently been proposed by different manufacturers and adopted in the field of packaging machines for their flexibility. For such systems, no prior instances of faulted bearings are available, and the size of a real fault is also unknown. Hand-made faults with drills did not produce discernible faults appreciable in post-processing of the data. Therefore, a picosecond laser with a pulse duration of 10 ps and a maximum energy per pulse of approximately 100 μJ is utilized to create a set of test bearings with increasing fault sizes on the outer race. Post-processing of the data enables the qualification of the minimum fault severity detectable in this specific application.

1. INTRODUCTION

The emerging technology of independent cart systems comprises a fleet of linear motors that operate on a shared track and can be controlled individually and independently. Unlike conventional motors, where the stator and rotor are enclosed within the same frame, linear motors feature a completely detached rotor. In this design, the fixed frame of the motor serves as the stator, with coils evenly distributed along the track, while the single cart functions as the movable part, containing permanent magnets positioned above the coils, along with a set of bearings. The cart, while crucial, is a completely passive component of the system. It interact with the changing magnetic field of the motor coils, moving in synchronization with the magnetic field pattern.

Independent cart systems stand poised to revolutionize traditional conveyor belts. Offering a range of benefits such as increased flexibility and dynamic capabilities, independent cart systems are proving to be ideal solutions for a wide range of industrial and motion control applications. Comprising modular components such as linear motors, guide rails, and control circuitry, these systems offer adaptability and efficiency. Linear motors within independent cart systems are

available in both straight and curved modules, allowing for diverse path configurations tailored to specific industrial requirements. Across industries, various iterations of independent cart systems are being deployed, each with distinct features in guide rail and cart design. Depending on application needs, guide rails may either attach directly to linear motors (Jabbar, D'Elia, & Cocconcelli, 2023; Jabbar, Cocconcelli, D'Elia, & Strozzi, 2023) or form a separate bed parallel to them (see figure 2), a brief description of such a system with parallel guide rail can also be found in (Cavalaglio Camargo Molano, Capelli, Rubini, Borghi, & Cocconcelli, 1968). Similarly, cart and bearings can take on different configurations to suit specific application requirements. For example, in a previous study (Jabbar, D'Elia, & Cocconcelli, 2023; Jabbar, Cocconcelli, et al., 2023), the authors have described setups where guide rails are directly attached to linear motors, with options for 12 or 6 bearings featuring plastic outer races. One significant advantage of independent cart systems is their ability to define the entire track in stations, enabling the accommodation of distinct speed profiles for each cart along every station.

Despite the considerable benefits that these systems offer over conventional conveyors, economic concerns remain. To date, the high cost and substantial upfront investment required for independent cart systems present challenges to their widespread adoption. To ensure economic viability, it is imperative that these systems operate flawlessly. Thus, the implementation of robust condition monitoring measures is crucial for their sustained performance. Although independent cart systems are rapidly replacing conveyor belt systems, there has been a lack of studies focused condition monitoring of such systems. Due to the confidentiality surrounding these systems, often protected by non-disclosure agreements, there is no repository or public database where vibration, acoustic, or other data related to independent cart systems is available for conducting condition monitoring. Consequently, to the best of the authors' knowledge, there has been no research conducted on condition monitoring of the independent cart systems. This scarcity of research and data necessary for condition monitoring serves as a primary motivation for this study. Furthermore, another motivation for this research stems from the inherent challenges associated with monitoring the condition of such systems.

Condition monitoring of the independent cart systems presents a formidable challenge for several reasons. Firstly, it is a highly non-synchronous system, with speed variations ranging from a few millimeters per second to several meters per second. For instance, the system under study can achieve cart's speed of up to 4 meters per second. Additionally, the system supports speed reversal, allowing for changes in direction. Furthermore, with the addition of every cart to the fleet, the number of bearings increases by three. Given that each cart contains three bearings and there are hundreds of bearings in the fleet, monitoring the condition of the independent cart systems becomes complex. This monitoring process can be divided into several steps: firstly, identifying if there is a fault; secondly, determining the type of fault (such as inner race, outer race, ball fault, etc.); thirdly, distinguishing whether the top bearing is faulty or the bottom one; and fourthly, localizing the cart carrying the faulty bearing. Understanding the behavior of the vibration signal in the case of a top bearing fault is also intriguing. This is because there is a pair of bearings with exactly the same dimensions and dynamics supporting the same movement.

There are several bearing datasets available, including the Intelligent Maintenance Systems (IMS) dataset (Lee, Qiu, Yu, & Lin, 2007), the Case Western Reserve University (CWRU) dataset (*Bearing Data Center, Case Western Reserve University (CWRU)*, n.d.), and the IEEE Dataport bearing dataset (B. Hu, 2023). These datasets serve as benchmarks for testing and evaluating new condition monitoring algorithms. However, these datasets primarily focus on conventional configurations where the bearings are fixed around a shaft. In contrast, in independent cart systems, the bearings not only rotate but also translate along the path defined by a set of linear motors. This introduces a significant departure from conventional setups. Additionally, the linear motion of the cart is often highly non-synchronous, resulting in non-synchronous rotational speeds of the bearings. Furthermore, the possibility of instant speed reversal further complicates the dynamics of the system. Given these differences, there is a clear need for new datasets specifically tailored for the condition monitoring of such industrial systems. These datasets would need to capture the unique characteristics and challenges posed by the operation of bearings within independent cart systems.

Data-driven modeling has emerged as a valuable approach that can complement traditional signal processing techniques, offering a robust methodology for uncovering insights from complex datasets. This methodology involves leveraging machine learning algorithms to discern patterns and anomalies within the data. While implementation requires expertise in both data preprocessing and machine learning, the benefits of this approach can be significant. In light of these advantages, researchers have explored the integration of machine learning algorithms for bearing fault classification, alongside classical signal processing techniques. A diverse array of algorithms has been deployed for this purpose, encompassing Long Short Term Memory (LSTM)((Walther & Fuerst, 2022), (Y. Hu et al., 2022)), Autoencoders ((K. Cheng RC.and Chen, 2022), (R.-C. Cheng, Chen, Liu, Chang, & Tsai, 2021)), Support Vector Machines (SVM) ((Sun & Liu, 2023), (Y. Fan, Zhang, Xue, Wang, & Gu, 2020)), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Deep Learning (DL), Transfer Learning (Wan et al., 2022), Anomaly Detection (Z. Fan et al., 2023), and the K-means clustering algorithm (Yiakopoulos, Gryllias, & Antoniadis, 2011), among others.

Although this list is not exhaustive, it underscores the breadth of approaches available for analysis. By combining these various methodologies, researchers aim to construct more precise and reliable models for machine condition monitoring. Through this integration, the potential for accurately identifying and predicting faults in machinery is significantly enhanced.

Machine learning-based condition monitoring of mechanical systems relies heavily on data. To amass a large dataset, it's essential to gather a vast array of data encompassing various degrees of faults and experimental conditions. Achieving this requires the creation of artificial faults through specific methods. Artificially injecting faults poses challenges, particularly when it comes to ensuring repeatability and control over the size of the fault. Traditional methods such as cutting tools or drill mills may lack accuracy, and be limited by the high hardness of the materials employed for bearings, in particular, when attempting to damage different parts of the bearing with varying fault sizes. Consequently, manually employing these tools often results in irregular fault sizes and dimensions. In contrast, laser injection offers a solution that provides repeatability and precise control over fault dimensions. By utilizing laser technology, faults can be consistently created even in very high strength material with accurate control over their size and shape, ensuring greater consistency and reliability in experimental conditions.

The rest of the paper is structured as follows. Section 2 provides insights into the experimental setup of the independent cart system, including the cart and bearing configuration. Following this, Section 3 elaborates on the setup involving picosecond laser technology. In Section 4, the paper delves into the specifics of the experimental campaign. Moving forward, Section 5 presents preliminary results derived from the conducted experiments. Finally, conclusions drawn from the study are discussed in Section 6.

2. Independent Cart System and Experimental Setup

The experimental setup for the independent cart system utilized in this study comprises eight straight motor modules, two curved modules, and a total of 12 carts. Each straight motor module features a 250-millimeter (mm) long stator, while the stator of the curved module measures 500 mm in length. Consequently, the combined length of the track defined by the 10 motor modules totals 3000 mm (see figure 1). With the system's capability to program each cart independently for a desired travel path, provided there are no collisions between carts, experiments were conducted with varying numbers of carts. Each cart, measuring 50 mm in length, consists of five pairs of permanent magnets and is equipped with a set of three rolling-element bearings as shown in figures 3 and 4. The two top bearings share identical geometry, with an external diameter of 25 mm, while the bottom bearing is slightly larger, with an external diameter of 35 mm. These bearings ensure secure attachment of the cart to the guide rail, which runs parallel to the linear motor modules (see figure 2). The motor modules are outfitted with coils that generate magnetic field patterns, enabling the system to achieve the desired trajectory and motion profile for each cart.

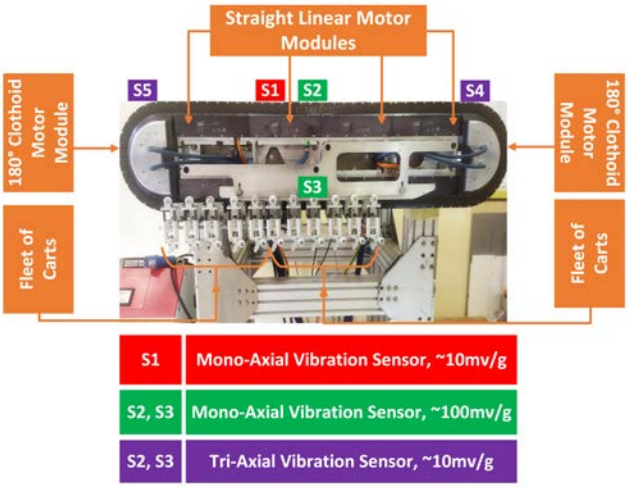


Figure 1. The independent cart system with Parallel guide rail.

To acquire vibration data, five vibration sensors were strategically installed at various locations throughout the system's geometry, as depicted in the figure 1. Among these sensors, three are monaxial, each with different sensitivities (e.g., 10mv/g and 100mv/g). The remaining two sensors are triaxial and are directly mounted onto the guide rail near the left and right sides of the system, with each axis exhibiting a sensitivity of approximately 10mv/g.

3. Picosecond Laser

An EKSPLA Atlantic 50 picosecond laser source was used to damage the bearings. It's a picosecond laser source which generates a Gaussian beam profile at the IR wavelength of 1064 nm. This kind of laser source produces ultrashort pulses in the picosecond regime thus allowing to ablate the inner (and the outer) race by creating a very precise in shape grooves without limitations in terms of hardness of the material to be harmed. The laser beam diameter in the focal position settled at $\phi \approx 10$ μm, evaluated at $1/e^2$ intensity. A dedicated optical path for the IR wavelength allowed the laser beam exiting from its source to be delivered to the scanning head. This is a Raylase Supercan IV galvanometric scanner copuled with an 80 mm F-theta lens thus allowing a square working area with a side of 39 mm. The remaining movements, outside the above scan area, were instead guaranteed by the translation, in X and in Y- direction, of a stage on which the bearings to be damaged were placed. The translation of Z axis

Figure 2. The parallel guide rail.

allowed to damage the bearings by working at the correct focal height. The entire system set up in BrightLab laboratory of the DISMI Department is shown in Figure 5. Preliminary tests were conducted to evaluate the response of the material to the infrared radiation. These tests allowed us to correctly identify, and define, all the process parameters to be able to carry out damage with specific geometry, dimension and prescribed depth. The adopted laser parameters are summarized in Table 1.

Table 1. Laser parameters used during experiments.

Parameter	Unit of Measure	Value
Wavelength, λ	nm	1064
Average Output Power, P	W	13.32
Pulse Frequency, f	kHz	300
Pulse energy, E	µJ	44.4
Pulse duration, τ	ps	10
Pulse fluence, F	J/cm^2	56.56
Line spacing, s	µm	5
Marking speed, v_s	m/s	1
Number of passes on each groove, p		150

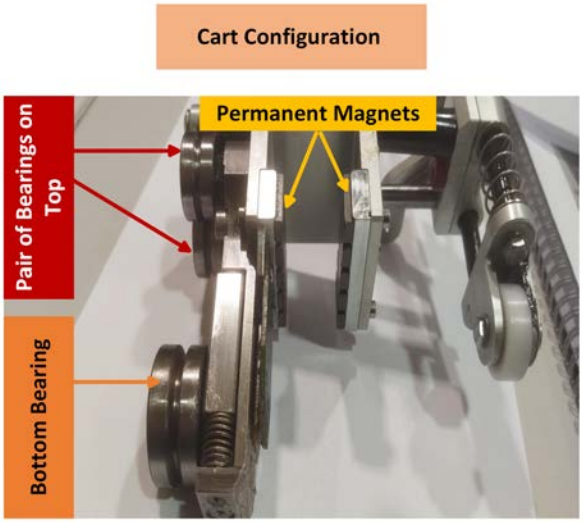


Figure 3. System's cart configuration.

4. Experimental Campaign

The experimental campaign can be divided into two main categories: fault injection and vibration data acquisition, with and without the presence of bearing faults. Fault injection campaigns, furthermore, can be categorized into two main types: those that do not involve dismantling the bearings and those that involve disassembled bearings. Initially, our approach aimed to create faults in bearings without the need for dismantling, utilizing Lab-available drill mill. However, due to the limited flexibility in maneuvering of the drill mill head caused by the close proximity of the bearing's inner and outer races, the faults injected were predominantly of an incipient nature. Consequently, these faults failed to fully encompass the entire grooves of the inner and outer races. Henceforth, these faults will be referred to as incipient fault types throught the remainder of this paper. Subsequently, we explored an alternative method using the EKSPLA Atlantic 50 picosecond laser to create faults in the bearings without disassembly. This approach allowed to reach controlled fault while entailing a challenging method to perform the damage of the bearing parts. The main difficulties were found in the creation of damage both in the inner race and in the outer race of the bearing and they referred to the following two aspects. The first is linked to the shielding effect of the laser radiation that some components of the non-disassembled bearing may have on the area that needs to be damaged. This inconvenience also occurred when specific supports were used to correctly orient (and strongly tilt) the bearing (e.g. the internal race didn't always allow the defect to be created in the external race in its entire axial extension due to the shielding effect performed). Instead, the second considered the need to create a groove with a predetermined depth on a strongly inclined component due to the considerations previously anticipated. To overcome these limitations, and also to be robust in ob-

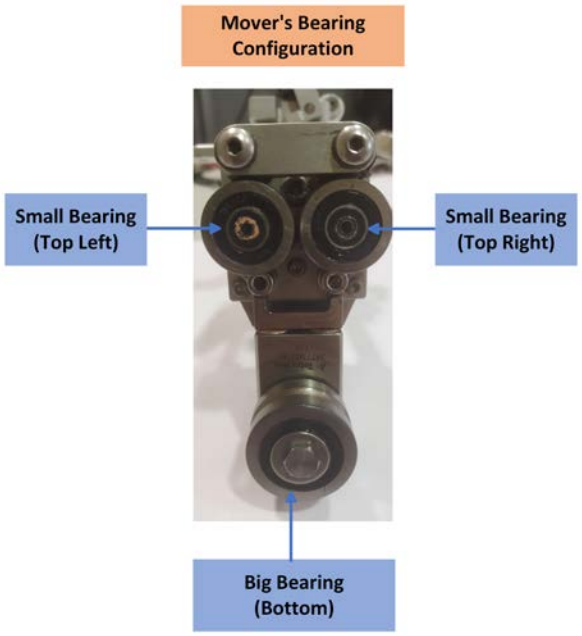


Figure 4. Cart's bearing configuration.

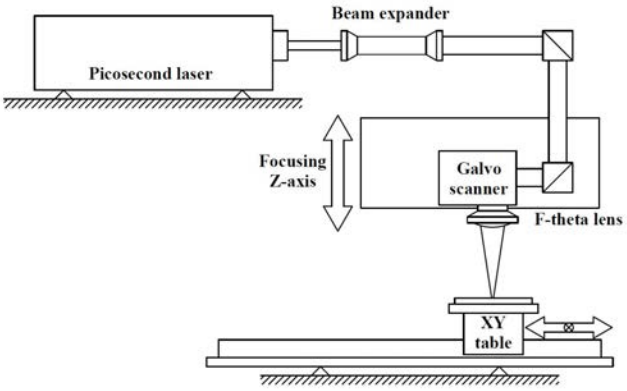


Figure 5. A schematic of the laser ablation system.

taining the damage on the various component of the bearings comparable to what was achieved on preliminary planar tests, an evaluation of the allowable depth of focus was required. In particular, by considering a beam quality factor M^2 of 1.5 and a laser beam diameter (evaluated at $1/e^2$ intensity) on the lens of $D_0 = 14$ mm, have been calculated:

- Rayleigh length z_f, as the distance from the beam waist where the beam radius is increased by a factor of the $\sqrt{2}$:

$$z_f = \frac{\pi \left(\frac{D_0}{2}\right)^2}{M^2 \lambda} \approx 100\ \mu m \tag{1}$$

- Depth of Field (D.O.F.), as the distance either side of the beam waist, D_0, over which the beam diameter grows by 5%

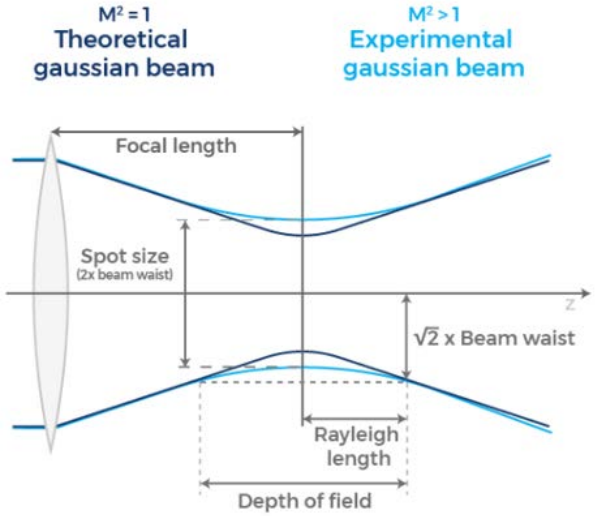


Figure 6. Pico-second laser beam profile

$$\text{D.O.F.} = \pm 0.08 \pi D_0^2 M^2 \lambda \approx 118\ \mu m \tag{2}$$

The modest value of the depth of Field (D.O.F.), which settles at 118 μm for our laser, did not allow for damaging the bearing components in a single laser pass. Therefore, a multi-pass approach was needed. This methodology involves focusing the laser beam on a target zone of the component that needs to be damaged. The process carried out is characterized by reaching the prescribed depth in the area where the laser was focused, and by unworked areas where the laser was unable to completely deposit its energy and properly ablate the material. The processing area can therefore be limited to a rectangle having as its height the depth of field previously calculated. The areas not included within this height will instead remain untreated due to the unfocused conditions. An incremental increase in depth is needed and allowed by varying the height of the galvanometric head (and correspondingly recovering the *X* position by advancing with the *XY* table). In this way, by focusing on a lower area, further ablation processing is possible. The process is performed continuously until the intended damage shape is achieved. The reached depth had a tolerance of ± 0.15 mm from the nominal shape. Finally, to fully overcome the limitations related to the first issue explained above, the bearings were completely dismantled. This way, individual parts of the bearing were fully exposed, facilitating controlled fault injection. The damage process was unaffected by the presence of, for instance, balls between the inner and outer races, which obscured the area to be damaged. Thus, better and more precise flat-bottom grooves were created using the multi-pass criteria mentioned above. An optical microscope (Mod. Nikon LV100ND), was used to characterize the final shape of the fault in terms of dimensions and, by changing the depth of focus.

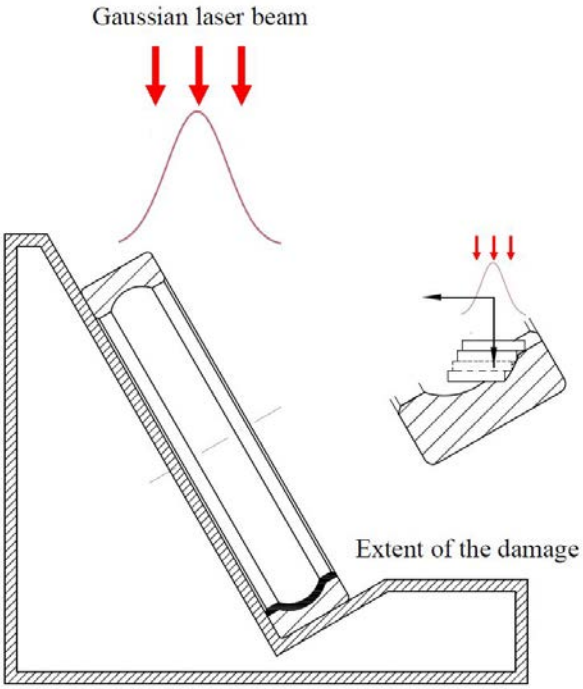


Figure 7. Positioning of bearings on an ad-hoc support. The operating principle of the multi-altitude approach for progressive focusing is explained in the enlarged image.

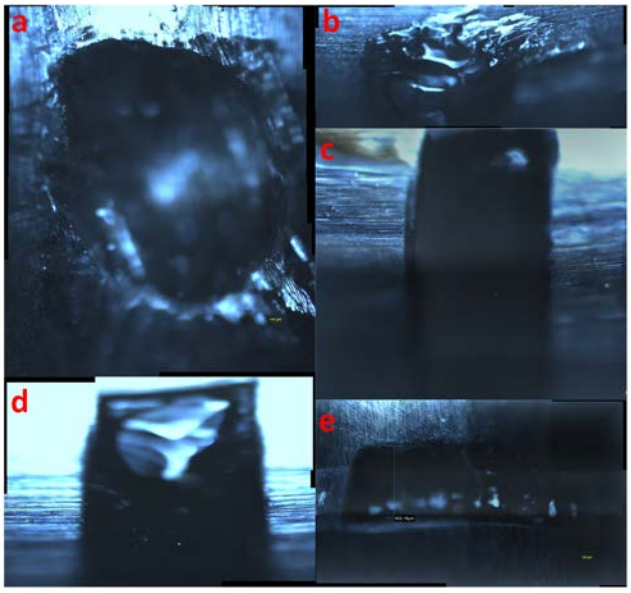


Figure 8. Microscopic view of bearing faults: a) Incipient version IR fault of bottom bearing, b) Incipient version OR fault of bottom bearing, c) Lase version 0.5 OR fault of top bearing, d) Lase version 1.0 OR fault of bottom bearing, e) Lase version 0.5 IR fault of bottom bearing.

Table 2. Bearing type, fault type, and fault version.

Bearing Type	Fault Type	Fault Version			
		Incipient Width	Laser 0.5 Width	Laser 1.0 Width	Laser 2.0 Width
Top Bearing	OR	NA	0.5	1.0	2.0
	IR	0.75	0.5	1.0	2.0
	BF	NA	0.5	1.0	2.0
	OROS	NA	0.5	NA	NA
Bottom Bearing	OR	1.3	0.5	1.0	2.0
	IR	1.4	0.5	1.0	2.0
	BF	NA	0.5	1.0	NA
	OROS	NA	0.5	NA	NA

The microscopic view of the some of the faults created without dismantling the bearings is illustrated in the figure 8. It is evident that the depth of the faults varies non-uniformly due to the challenges mentioned earlier. A tabular summary of the fault injection campaign without dismantling the bearings is presented in the table 2. In the table, ”IR” represents inner race faults, ”OR” denotes outer race faults, ”BF” signifies ball/roller faults, and ”OROS” indicates outer race outer surface faults. Whereas, laser versions 0.5, 1.0, and 2.0 represent the nominal width (in mm) of the injected faults.

5. Results

The experiments involved varying the number of movers and different experiment types, each focusing on different areas of the system, such as the straight path on the top or bottom side, or the curved side, with varying numbers of movers. However, the results presented are specifically for a particular experiment type involving a single mover traversing the straight path on the top side of the system, demonstrating back-and-forth motion. This experiment was conducted at different speeds: 1000 mm/s, 2000 mm/s, and 3000 mm/s. It is important to note that the results will not be discussed in terms of fault identification based on fault frequency. Instead, the focus will be on highlighting differences in the vibration data concerning fault injection types and fault sizes.

The figure 9 illustrates the Fast Fourier Transform (FFT) of the data obtained during the experiment conducted at a speed of 1000mm/s, showcasing different versions of IR faults in the top bearing. In the legend, ”F” indicates the presence of a fault, while ”H” denotes healthy conditions. Additionally, ”S1000” represents the linear speed of the cart. ”V” signifies the laser version of fault injection, while ”Incipient Fault” denotes manual fault injection. It is important to note that only one of the top couple of bearings is faulty at a time. Moreover, the width of the impulses in the data with faulty conditions appears to be directly correlated with the size of the faults; sharper faults result in narrower impulse widths. The same behavior is observed in the case of the outer race fault of the bottom bearing, as illustrated in the figure 12.

As evident from the table 2, the Incipient type IR fault in the top bearing measures 0.65mm, which closely aligns with the laser version 0.5 fault. This similarity is also reflected

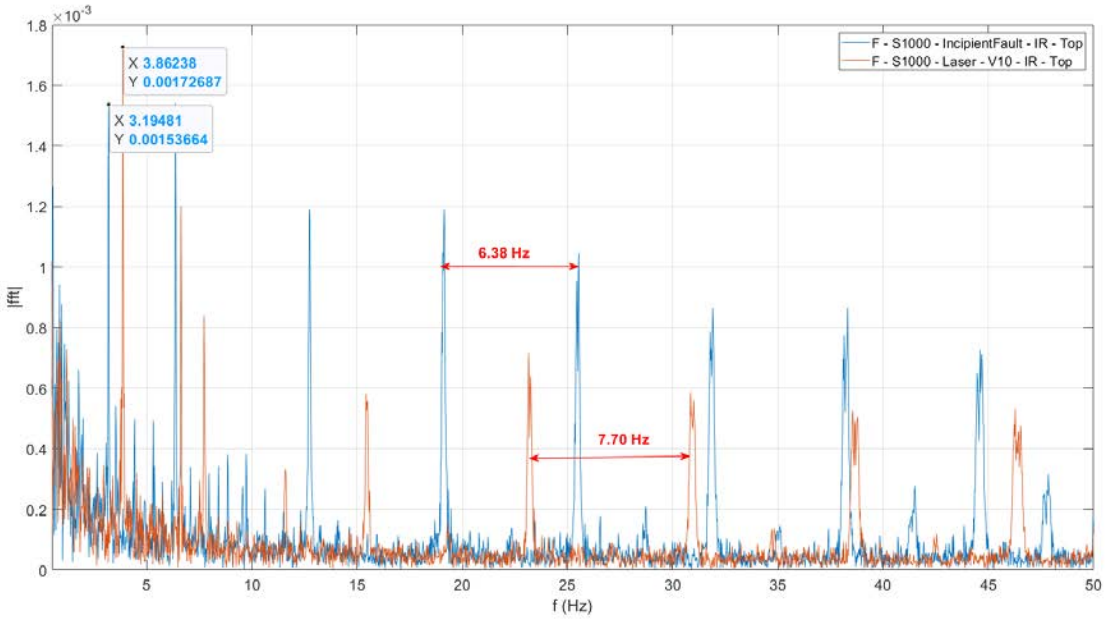


Figure 9. The FFT of the vibration data with an incipient and laser version 1.0 type inner race fault in the top bearing, at a linear speed of 1000 mm/s.

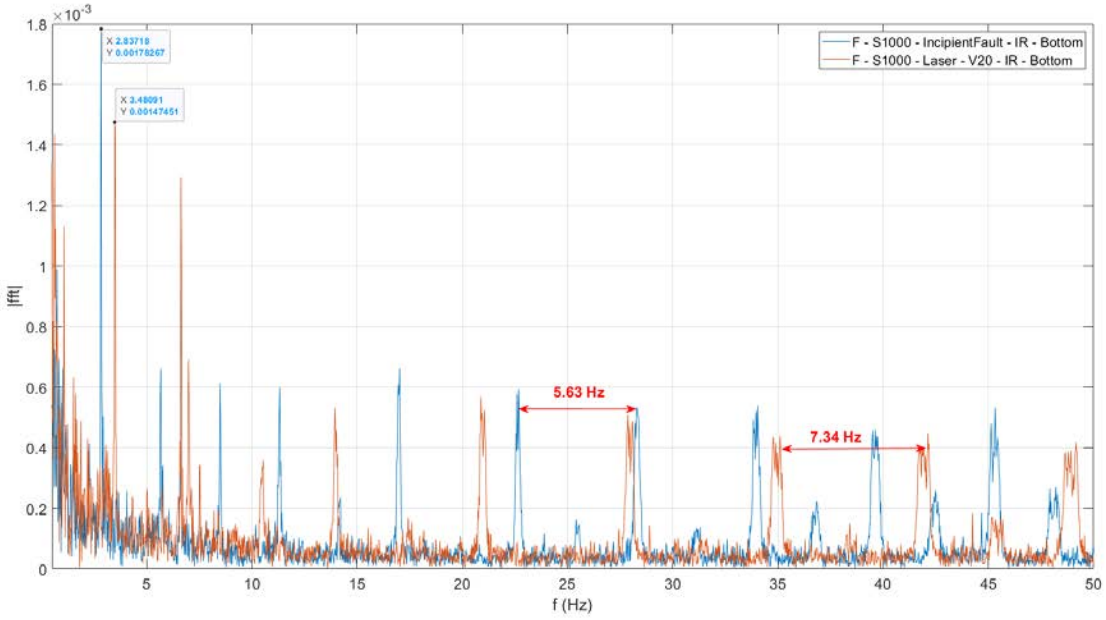


Figure 10. The FFT of the vibration data with an incipient and laser version 2.0 type inner race fault in the bottom bearing, at a linear speed of 1000 mm/s.

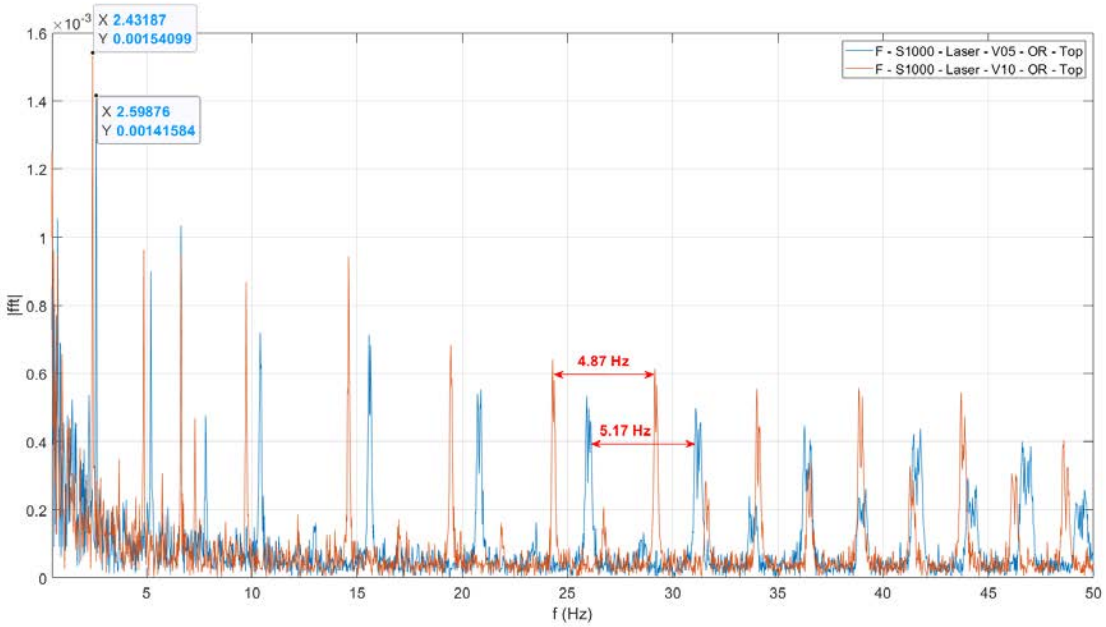


Figure 11. The FFT of the vibration data with laser version 0.5 and laser version 1.0 type outer race faults in the top bearing, at a linear speed of 1000 mm/s.

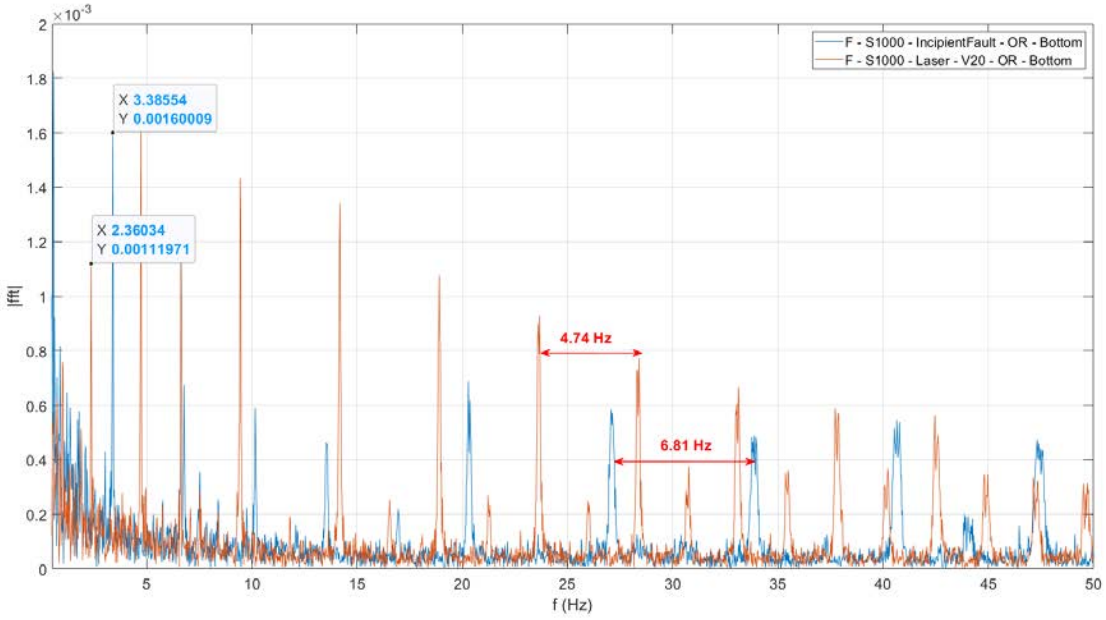


Figure 12. The FFT of the vibration data with an incipient and laser version 2.0 type outer race fault in the bottom bearing, at a linear speed of 1000 mm/s.

in the width of the impulses in the FFT of the vibration signal. However, it's noteworthy that the width of the incipient fault is not uniform throughout the fault area, nor does it completely cover the groove of the inner race where the rolling element glides. Therefore, a detailed explanation and analysis of these differences are yet to be conducted. Similarly, in the case of the IR bottom bearing fault, the incipient fault, approximately 1.4 mm wide at its widest, shows FFT characteristics more akin to the laser version 2.0 fault type as depicted in figure 10. Notably, for the incipient type IR fault of the bottom bearing, the modulating frequency appears to be 2.83 Hz, with even harmonics exhibiting higher energy than the odd harmonics. Conversely, for the laser version 2.0 fault in the bottom bearing, the modulating frequency seems to be 3.48 Hz, with even harmonics having more energy than the odd harmonics.

Likewise, for the IR top bearing fault in figure 9, the incipient type fault measuring approximately 0.75mm exhibits FFT characteristics more similar to the laser version 1.0 fault type. Interestingly, for the incipient type IR fault of the top bearing, the modulating frequency appears to be 3.19 Hz, with even harmonics having higher energy than the odd harmonics. Conversely, for the laser version 1.0 IR fault in the top bearing, the modulating frequency seems to be 3.86 Hz, with even harmonics exhibiting more energy than the odd harmonics.

Regarding the OR bottom bearing fault, the incipient fault, approximately 1.3 mm wide at its widest, exhibits FFT characteristics more aligned with the laser version 2.0 fault type as shown in figure 12. Notably, for the incipient type OR fault of the bottom bearing, the modulating frequency appears to be 3.38 Hz, with even harmonics demonstrating higher energy than the odd harmonics. Conversely, for the laser version 2.0 fault in the bottom bearing, the modulating frequency seems to be 2.36 Hz, with even harmonics having more energy than the odd harmonics.

Lastly, in the case of the OR top bearing fault in figure 11, the FFT of the laser version 0.5mm resembles more closely the laser version 1.0 fault type. Interestingly, for the version 0.5 OR fault of the top bearing, the modulating frequency appears to be 2.59 Hz, with even harmonics demonstrating higher energy than the odd harmonics. Conversely, for the laser version 1.0 fault in the top bearing, the modulating frequency seems to be 2.43 Hz, with even harmonics exhibiting more energy than the odd harmonics.

6. Conclusion

In order for the dataset to be utilized in developing new algorithms, it is imperative that at least a subset of the data exhibits clear fault signatures. The laser method of fault injection offers precise control over fault dimensions, ensuring repeatability. The data acquired from this experimental campaign could serve as a benchmark for the development and testing of condition monitoring algorithms, whether they are machine learning-based, statistically based, or employ classical signal processing techniques.

Acknowledgment

Authors gratefully acknowledge the European Commission for its support of the Marie Sklodowska Curie Program through the H2020 ETN MOIRA project (GA 955681).

References

Bearing data center, case western reserve university (cwru). (n.d.). Retrieved from {http://csegroups.case.edu/bearingdatacenter/hom.}

Cavalaglio Camargo Molano, J., Capelli, L., Rubini, R., Borghi, D., & Cocconcelli, M. (1968). Determination of threshold failure levels of semiconductor diodes and transistors due to pulse voltages. *IEEE Transactions on Nuclear Science*, *15*(6), 244-259.

Cheng, K., RC.and Chen. (2022). Ball bearing multiple failure diagnosis using feature-selected autoencoder model. *Int J Adv Manuf Technol 120, 4803–4819 (2022). https://doi.org/10.1007/s00170-022-09054-x*(120), 4803–4819.

Cheng, R.-C., Chen, K.-S., Liu, Y., Chang, L.-K., & Tsai, M. C. (2021). Development of autoencoder-based status diagnosis method for ball bearing tribology status monitoring. *The Proceedings of The 9th IIAE International Conference on Industrial Application Engineering 2020*. Retrieved from https://api.semanticscholar.org/CorpusID:236644837

Fan, Y., Zhang, C., Xue, Y., Wang, J., & Gu, F. (2020). A bearing fault diagnosis using a support vector machine optimised by the self-regulating particle swarm. *Shock and Vibration*. Retrieved from https://api.semanticscholar.org/CorpusID:214661891

Fan, Z., Wang, Y., Meng, L., Zhang, G., Qin, Y., & Tang, B. (2023). Unsupervised anomaly detection method for bearing based on vae-gan and time-series data correlation enhancement (june 2023). *IEEE Sensors Journal*, *23*(23), 29345-29356. doi: 10.1109/JSEN.2023.3326335

Hu, B. (2023). *bearing dataset.* IEEE Dataport. Retrieved from https://dx.doi.org/10.21227/5p7e-pz02 doi: 10.21227/5p7e-pz02

Hu, Y., Wei, R., Yang, Y., Li, X., Huang, Z., Liu, C., Y.and He, & Lu, H. (2022). Performance degradation prediction using lstm with optimized parameters. *Sensors 2022, 22, 2407. https://doi.org/10.3390/s22062407*, *2407*(22).

Jabbar, A., Cocconcelli, M., D'Elia, G., & Strozzi, R., M.and Rubini. (2023). Results on experimental data analysis of independent cart systems in non-stationary conditions. In *Surveillance, vibrations, shock and noise, institut supérieur de l'aéronautique et de l'espace [isae-supaero], jul 2023, toulouse, france.*

Jabbar, A., D'Elia, G., & Cocconcelli, M. (2023). Experimental setup for non-stationary condition monitoring of independent cart systems. In *In: Kumar, u., karim, r., galar, d. and kour, r. (eds). international congress and workshop on industrial ai and emaintenance 2023. iai 2023. lecture notes in mechanical engineering.*

Lee, J., Qiu, H., Yu, G., & Lin, J. (2007). *Rexnord technical services, ims, university of cincinnati. "bearing data set", nasa ames prognostics data repository, nasa ames research center, moffett field, ca.* Retrieved from {http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/.}

Sun, B., & Liu, X. (2023). Significance support vector machine for high-speed train bearing fault diagnosis. *IEEE Sensors Journal*, *23*(5), 4638-4646. doi: 10.1109/JSEN.2021.3136675

Walther, S., & Fuerst, A. (2022). Reduced data volumes through hybrid machine learning compared to conventional machine learning demonstrated on bearing fault classification. *Appl. Sci. 2022, 12, 2287. https://doi.org/10.3390/app12052287*, *2287*(12).

Wan, S., Liu, J., Li, X., Zhang, Y., Yan, K., & Hong, J. (2022). Transfer-learning-based bearing fault diagnosis between different machines: A multi-level adaptation network based on layered decoding and attention mechanism. *Measurement*, *203*, 111996. Retrieved from https://www.sciencedirect.com/science/article/pii/S0263224122011927 doi: https://doi.org/10.1016/j.measurement.2022.111996

Yiakopoulos, C., Gryllias, K., & Antoniadis, I. (2011). Rolling element bearing fault detection in industrial environments based on a k-means clustering approach. *Expert Systems with Applications*, *38*(3), 2888-2911. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417410008791 doi: https://doi.org/10.1016/j.eswa.2010.08.083

Uncertainty in Aircraft Turbofan Engine Prognostics on the C-MAPSS Dataset

Mariana Salinas-Camus[1] and Nick Eleftheroglou[2]

[1, 2] *Intelligent Sustainable Prognostics Group, Aerospace Structures and Materials Department, Faculty of Aerospace Engineering, Delft University of Technology*
m.salinascamus@tudelft.nl
n.eleftheroglou@tudelft.nl

ABSTRACT

Prognostics and Health Management (PHM) plays a crucial role in maximizing operational efficiency, minimizing maintenance costs, and enhancing system reliability. Predicting Remaining Useful Life (RUL) is a key aspect of PHM, inherently incorporating uncertainty. This paper focuses on uncertainty quantification (UQ) within Data-Driven Models (DDMs), particularly Machine Learning (ML), such as Long Short-Term Memory (LSTMs), and stochastic models namely Hidden Markov Models (HMMs). While ML models emphasize accuracy, stochastic models offer a different paradigm for prognostics, directly addressing uncertainty. Traditional categorizations of uncertainty as aleatory and epistemic face challenges in practical implementation. This paper explores how, in prognostics, HMMs primarily tackle aleatory uncertainty, whereas LSTMs predominantly address epistemic uncertainty. It also discusses the complexities of uncertainty management in prognostics and analyzes further an already proposed alternative approach to categorize uncertainties. Despite theoretical advancements, practical implementation remains challenging, especially for DL models due to their limited interpretability. This study sheds light on UQ challenges and offers insights for future research directions in prognostics.

1. INTRODUCTION

Prognostics and Health Management (PHM) is a field that provides users with a thorough analysis of both the current and future health condition of a system. PHM has gained attention during the last years due to the potential that it has to maximize the operational availability, reduce maintenance costs, and improve the system reliability.

Prognostics, as part of the PHM field, aim at predicting the Remaining Useful Life (RUL) of a given engineering system while it is in operation. By definition, the prediction of RUL incorporates uncertainty. Therefore, it is imperative to model RUL as a random variable rather than a deterministic one to account for the inherent uncertainties in prognostics. The prediction of RUL is then used by a decision-making module, which will make health management decisions to fulfill PHM goals.

Nonetheless, uncertainty quantification (UQ) is a challenge within prognostics. In particular, when prognostics are performed with Data-Driven Models (DDMs), which only rely on historical sensor data, UQ can become a greater challenge depending on which type of prognostic DDM is used. Hence, this paper will solely focus on UQ for DDMs, given their extensive use in prognostics and their sensitivity to uncertainty sources related to the data.

As previously mentioned, DDMs use historical sensor data to predict the RUL of the engineering system, and there are different types of DDMs. For the purposes of this paper, we would consider Machine Learning (ML) and stochastic models as the two main categories of DDMs. ML models, which include decision trees, Support Vector Regressor (SVR), and Deep Learning (DL) models, among others, have gained attention in prognostics because of the high accuracy of the RUL predictions.

In contrast to ML, stochastic models offer a different paradigm for data-driven prognostics. Stochastic models, such as Hidden Markov Models (HMMs) and Wiener processes, model the degradation process of the engineering system based on the sensor data, i.e. unsupervised learning, unlike ML models that can find complex relationships between the sensor data and RUL, i.e. supervised learning. Thus, ML models find patterns in the data which allows them to have a good performance when trained with large and labeled datasets but struggle with outliers.

Another important difference between ML models and

stochastic models is UQ. In stochastic models, UQ is direct since the output of the model is a probability density function (pdf) of the RUL prediction. For ML models, UQ is a challenge given that ML models are deterministic by nature, i.e. they provide a single-point prediction for RUL. There are techniques to perform UQ, but these might not be suitable for a prognostic application, as will be discussed in Section 3.

It is important to know that UQ is not the ultimate goal, but it is a key step towards uncertainty management. Uncertainty management is defined as the identification of sources of uncertainty and the reduction of uncertainty by leveraging data to characterize better the inherent prognostic uncertainties. Hence, reducing their impact on RUL predictions, which is necessary for the decision-making process (Sankararaman, 2015).

The question is then, which are the sources of uncertainty? The classical categorization considers two sources of uncertainty, aleatory and epistemic. Aleatory refers to the uncertainties that are intrinsic randomness of a phenomenon. Epistemic uncertainty is caused by a lack of knowledge, thus, it is the uncertainty that comes from the model itself (Der Kiureghian & Ditlevsen, 2009). Another way to look at them is that aleatory uncertainty is irreducible, since there is no control over the randomness of the phenomenon, and epistemic uncertainty is reducible given that the model can be changed. Consequently, to perform uncertainty management epistemic uncertainty needs to be addressed.

Nevertheless, even if we manage to identify epistemic uncertainty effectively, how can this information be used to perform uncertainty management in prognostics? Aside from the variability of the data, i.e. aleatory uncertainty, there is uncertainty in the identification of the current state of the system's health or the future loading operation that the system will be subjected to. By considering all these different sources as part of "epistemic uncertainty", it is unclear what actions need to be taken to reduce the RUL uncertainty.

For that reason, it has been claimed that the aleatory and epistemic categorization is not suitable for prognostics (Sankararaman & Goebel, 2013) and a more suitable categorization has been proposed, which will be further explained in Section 5. Although this categorization has been presented in different publications, it has not been applied, to the best of the author's knowledge, to a real-life scenario. Until now, the few prognostics publications that identify sources of uncertainty continue to use the classical categorization.

This paper presents both an stochastic model and a DL model under the same case study. To understand the use of stochastic models, an HMM presented. With the HMM, a new expression for RUL prediction is introduced in this study and is compared with the state-of-the-art RUL expression in terms of UQ. For DL models, a Long-Short Term Memory (LSTM) is used, given that it has been argued as the one with the best performance in terms of accuracy for several engineering applications. The LSTM is analyzed by using different parameters for UQ.

Therefore, by the use of these models this paper aims to provide an understanding of uncertainty in prognostics, and how different types of DDMs deal with UQ. As well as to offer a discussion in terms of future perspectives to address the UQ challenge, ultimately aiming towards the goal of uncertainty management.

The paper is organized as follows, Section 2 offers the theoretical background of HMMs and the new prognostic expression and Section 3 details the UQ methods for DL models, as well as DL model approaches in prognostics. The case study, including the data prepossessing and results, is presented in Section 4. Section 5 offers a discussion about the future perspective on UQ for prognostics. Finally, the paper is concluded in Section 6.

2. HIDDEN MARKOV MODELS

HMMs are a widely used stochastic model for different engineering applications. In the context of prognosis, it has been used for composites (Eleftheroglou, 2020; Eleftheroglou et al., 2024), lithium-polymer batteries (Eleftheroglou et al., 2019), turbofan engines (Giantomassi et al., 2011), and simulated fatigue crack growth (Le et al., 2014). In each one of these publications, different variants of HMM are used. A multi-branch HMM is used in (Le et al., 2014) to take into account the multiple degradation modes that can occur. A more complex version of HMM is used in (Eleftheroglou, 2020; Eleftheroglou et al., 2019), called the Non-Homogeneous Hidden Semi Markov Model (NHHSMM). When applied to composites the author used an adaptive approach of the NHHSMM that allowed the model to predict the RUL for testing data that had gone through unexpected phenomena.

Thus, HMMs have demonstrated their applicability through the use of different variants of it. In this paper, the classical HMM will be used, and a new definition for prognostic is presented.

HMMs can model a sequence of observations, which in this case is the data coming from the sensors. It is used in processes in which the state of the engineering system cannot be directly observed, hence, they are hidden. The engineering system is modeled as a Markov process, meaning that the probability of transitioning from one state to another depends only on the current state. The sojourn time of each hidden state is defined by an exponential distribution (continuous case) or a geometrical distribution (discrete case). Each state emits an observation with a certain probability distribu-

tion. Below, the parameters that describe an HMM are detailed (Rabiner, 1989).

- N: number of states. Individual states are denoted as $S = \{S_1, S_2, ..., S_N\}$, and the state at time t as q_t.
- M: number of distinct observation symbols per state. Individual observations are denoted as $V = \{v_1, v_2, ..., v_M\}$.
- State transition: the state transition probability distribution is denoted as $A = \{a_{ij}\}$, where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$. This expression is the probability that the state at time $t + 1$ is equal to the hidden state S_j given that the current state q_t is equal to the hidden state S_i.
- Observation distribution: the observation symbol probability distribution in state j, $B = b_j(k)$, where $b_j = P[v_k \text{ at t} | q_t = S_j]$, with $1 \leq j \leq N$ and $1 \leq k \leq M$.
- Initial state: the initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P[q_1 = S_i]$ with $1 \leq i \leq N$.

The complete parameter set of the model is denoted as $\lambda = (A, B, \pi)$. To train an HMM it is necessary then to adjust the model parameters λ to maximize $P(O|\lambda)$, meaning that the parameters are optimized to best describe the observation sequences, which in the case of prognostics are the degradation histories. Since there is no possible way of analytically calculating $P(O|\lambda)$, the iterative algorithm Baum-Welch can locally maximize it.

In the particular case of prognostics and this paper, some assumptions are made. First, the last state is not hidden but observable and it represents failure. Second, in the failure state, only one observation value is emitted. Third, only left-to-right transitions are allowed, meaning that while in hidden state i, it is only possible to remain in state i or to transit to state $i + 1$. This last assumption is valid only when modeling a degradation process independently from maintenance actions.

Once the model parameters λ are estimated and we have an observation sequence $O = O_1 O_2 ... O_T$, two questions arise. First, what is the probability of the observation sequence given the model $P(O|\lambda)$? The second question is, which is the most likely sequence of hidden states $Q = q_1 q_2 ... q_T$?

The answer to the first question, the Forward-Backward algorithm is used. The forward part calculates the likelihood of being in a hidden state at a certain time point given the available observations. The result of the forward part is then $P(O|\lambda)$, which answers the first question. The backward part is then used to answer the second question since calculates the likelihood of observing the remaining data, given the current hidden state. The result of the complete Forward-Backward algorithm is the posterior distribution which is the probability of being in each state at each time, given the entire sequence of observed data.

However, to answer fully to the second question it is necessary to find the single best state sequence that maximizes $P(Q|O, \lambda)$ that is equivalent to maximizing $P(Q, O|\lambda)$. The maximization is done via the Viterbi algorithm. After the Viterbi has estimated the most likely sequence of hidden states, it is possible to calculate the RUL. In the state-of-the-art, a time-invariant (TI) (Dong & He, 2007a) prognostic measure used is defined in (1). The variables $a_{i,i}$ and $a_{i,i+1}$ represent the probability of remaining in the current hidden state or transitioning to the next hidden state, respectively. The variable $D_i(d)$ represents the pdf (or pmf for the discrete case) evaluated in the probability of transition to the same state i, i.e. a_{ii}.

$$RUL_i^t = a_{i,i}(D_i(d) + RUL_{i+1}) + a_{i,i+1}(RUL_{i+1}) \quad (1)$$

In this paper, a new time-dependent (TD) prognostic measure is introduced in (2). This TD prognostic measure is expected to improve accuracy of the RUL prediction and to reduce the spread of the confidence intervals, which can be calculated by the weighted spread of uncertainty (WSU) presented in Appendix A.

$$RUL_i^t = d_{i,i}^T \left(D_i(d - \tau) + \sum_{i=k+1}^{N-1} D_k(d) + \mathcal{N}(1, \epsilon) \right) + d_{i,i+1}^T \left(\sum_{k=i+1}^{N-1} D_k(d) + \mathcal{N}(1, \epsilon) \right) \quad (2)$$

The notation for this expression is as follows. RUL_i^t, is the RUL in the state i and time step t. Once again, $D_i(d)$ represents the pdf (or pmf for the discrete case) evaluated in the probability of transition to the same state i, i.e. a_{ii}. The variable τ is the time spent in the current state i. Therefore, the term $D_i(d - \tau)$ represents a shift in the pdf making this RUL expression time-dependent. The variables $d_{i,i+1}^T$ and $d_{i,i}^T$ are derived from the transition matrix and are defined as shown in (3) and (4), respectively.

$$d_{i,i+1}^T = P(d \leq \tau | S_t = i) \quad (3)$$

$$d_{i,i}^T = 1 - d_{i,i+1}^T \quad (4)$$

The result of the expression 2 is the pdf of RUL per time step. Therefore, the confidence intervals can easily be obtained by calculating the cumulative density function (CDF) and, later, choosing the confidence level, usually 95%.

However, even if the HMM has a closed form for the posterior distribution, the distribution captures aleatory uncertainty, including uncertainty propagation and quantification via the prognostic measure. In state-of-the-art prognostics,

including the publications mentioned above, the HMMs presented usually address only aleatory uncertainty. Yet, epistemic uncertainty can be included in HMMs through a time-consuming sensitivity analysis, which traditionally has been used for accounting for epistemic uncertainty in stochastic models. In (Xie et al., 2016) a Generalized Hidden Markov Model (GHMM) is introduced that can identify both epistemic and aleatory uncertainties by using imprecise probabilities. The results show that the GHMM can make more robust decisions because the uncertainties can be differentiated. Yet, it is a computationally expensive model.

3. DEEP LEARNING MODELS

For DL models, as well as for any ML model, uncertainty quantification is a challenge since they are by nature deterministic, i.e. a single-point value for RUL prediction. Bayesian Neural Networks (BNNs), which are an extension of Neural Networks (NNs), overcome this by providing a pdf as a result. However, BNNs still have a problem when quantifying uncertainty since they offer an approximation of the posterior distribution (Abdar et al., 2021). The posterior distribution cannot be directly calculated because it is intractable to calculate the marginal distribution. Therefore, there is no close-form expression for the posterior distribution.

BNNs can provide a pdf as an output because they have distributions over the weights parameters and not deterministic values as in the case of NNs. These distributions in the weights parameters are learned over Bayesian inference, which uses the Bayes rules as shown in equation (5). In this expression, $P(w|X,Y)$ is the posterior distribution, $P(w)$ is the prior distribution, $P(X,Y|w)$ is the likelihood and $P(X,Y)$ is the marginal distribution.

$$P(w|X,Y) = \frac{P(X,Y|w)P(w)}{P(X,Y)} \tag{5}$$

Once again, it is computationally intractable to calculate $P(X,Y)$. Thus, these models offer an approximation of $P(w|X,Y)$ by using Variational Inference (VI). VI approximates the posterior distribution by using a variational parameter $q_\theta(w)$. The distribution q_θ is approximated by minimizing θ with the Kullback-Leibler (KL) divergence.

$$KL(q_\theta(w)||P(w|X,Y)) = \int q_\theta(w) \log \frac{q_\theta(w)}{P(w|X,Y)} dw \tag{6}$$

However, KL minimization is still intractable because it needs the posterior distribution that it was impossible to obtain in the first place. By rearranging KL into the evidence lower bound (ELBO), the need to have the posterior is avoided.

$$\mathcal{L}_{VI}(\theta) = \int q_\theta \log P(Y|X,w) dw - KL(q_\theta(w)||P(w)) \tag{7}$$

However, even though VI offers a good approximation of the posterior it is still challenging to implement given their computational cost (Nastos, Komninos, & Zarouchas, 2023). As a result, other techniques have arisen, such as Monte Carlo (MC) dropout, Deep Gaussian Processes, and Markov Chain Monte Carlo (Abdar et al., 2021).

MC Dropout has been introduced as a technique to quantify epistemic uncertainty and is the most used one due to its simple implementation (Gal & Ghahramani, 2016). This technique approximates the posterior by randomly switching off neurons, given a dropout probability. The same architecture is run multiple times and each dropout configuration corresponds to a different sample from the approximate posterior distribution.

However, MC dropout struggles to approximate complex posterior distributions, which may lead to good approximations only in certain regions of the posterior distribution but poor approximations in others (Fort, Hu, & Lakshminarayanan, 2019). Even more, it has even been questioned the fact that MC dropout is Bayesian since it fails sanity checks and is a design artifact since the posterior distribution converges to different values depending on the dropout probability assigned by a user (Folgoc et al., 2021). Hence, these techniques although easy to implement, do not always provide a good approximation of the desired distribution. The latter leads to uncertainty about the posterior distribution approximation that is already quantifying RUL uncertainty, adding up to uncertainty propagation of the entire prognostic model.

In the context of prognostics, these types of models have been applied with Bayesian LSTMs since LSTM, in general, provides the best results in terms of accuracy metrics. However, in (Peng, Ye, & Chen, 2019) and in (Xiahou, Wang, Liu, & Zhang, 2023) a point estimation of the final RUL value is made, instead of a prediction of RUL through the operation time. Nevertheless, (Xiahou et al., 2023) includes a RUL prediction during the operation time by including a credible interval. The results are promising, yet the main drawback of this approach is the complexity of the model and its optimization, as the authors have claimed to be "extremely intractable and time-consuming".

Other Bayesian approaches such as (Caceres, Gonzalez, Zhou, & Droguett, 2021) perform UQ including both aleatory and epistemic uncertainty. However, it is not reported in the results how much each source contributes to the confidence intervals, which are also quite volatile. Epistemic uncertainty is quantified with MC dropout with a probability dropout

value of 0.25, which is considered lower than the standard value of 0.5.

In (Pei et al., 2022) a Bayesian RNN is used with the dropout technique, however, they use a value of dropout between 0.05 to 0.2, which once again is considered low given that the standard dropout value. Low dropout values lead to narrow confidence intervals, meaning less estimated uncertainty in the RUL predictions. Thus, the choice of low dropout values can cause an underestimation of uncertainty that can be prejudicial for decision-making.

When it comes to aleatory uncertainty in DL models, it is split into two categories: homoscedastic and heteroscedastic. Homoscedastic uncertainty corresponds to the noise in the data and it remains constant through the whole data set, while heteroscedastic uncertainty corresponds to the noise that varies with the input (Nemani et al., 2023). The few DL models that include aleatory uncertainty, include only one part of it. For example in (Li, Yang, Lee, Wang, & Rong, 2020) a Bayesian DL framework is developed that takes into account heteroscedastic aleatory uncertainty. In the already mentioned work of (Caceres et al., 2021), only heteroscedastic aleatory uncertainty is address and it is also assumed to follow a Gaussian distribution.

4. Case Study

To perform a comparison between a stochastic and a DL model, the C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset is used (Frederick, DeCastro, & Litt, 2007). The C-MAPSS dataset is used as a benchmark within the prognostics community. This dataset is composed of four sub-datasets of simulated run-to-failure degradation histories from turbofan engines, with information from 21 sensors. Each sub-dataset considers a variety of operational conditions and injects different fault modes. For this paper only the sub-dataset FD001 is used, which consists of 100 training degradation histories. This dataset is divided into two in a random manner to have a training set of 80 degradation histories for training and 20 for testing. Additionally, sensors 1, 5, 6, 10, 16, 18, and 19 were eliminated from all the analyses since the values were fixed for every time measurement.

4.1. Pre-processing and training phase

For the HMM, only one feature can be used given the capabilities of the library used in Matlab. Therefore, sensor 11 is chosen since it is the sensor with the highest correlation to RUL. The sensor data is then discretized into 20 clusters using K-Means. The number of clusters was chosen based on the monotonicity index (MI), which allows to identification of the optimal number of clusters that can reasonably represent the degradation process. Once the data has been pre-processed, the optimal number of states is identified as 10, via the Bayesian Information Criterion (BIC). The expressions and results of both the MI and the BIC are shown in Appendix A, along with the estimated transition and emission matrices.

For the LSTM, first, an analysis of the importance of the sensors with respect to RUL was done. The sensors were selected based on their absolute Pearson Correlation Coefficient (PCC) with respect to RUL. Table 1 shows the results for all the sensors under analysis. The sensors selected were the ones with an absolute PCC higher than 0.6. Thus, sensors 2, 4, 7, 11, 12, 15, 17, 20, and 21 were used to train the LSTM. It is important to keep in mind that the LSTM is being trained with more data than the HMM, which only uses data coming from one sensor.

LSTMs need to receive sequences that have the same length, thus, the degradation histories were modified to fulfill this requirement. A sequence length of 362 was selected and values zeros were added in the RUL column, while for the sensors the last measurement was repeated. Thus, the shape of the training set tensor is (80, 362, 9).

The architecture of the LSTM is displayed in Figure 1. The last layer, which corresponds to a Dense layer, uses a linear function as activation. The model was trained using Adam as an optimizer, with a Mean Squared Error (MSE) loss function for 30 epochs.

Table 1. Sensor correlation to RUL values for dataset FD001 in C-MAPSS.

Sensor	PCC
2	-0.61
3	-0.58
4	-0.68
7	0.66
8	-0.56
9	-0.39
11	-0.7
12	0.67
13	-0.56
14	-0.31
15	-0.64
17	-0.61
20	0.63
21	0.64

4.2. Results and Discussion

The results are examined individually because the uncertainty captured by the HMM pertains to aleatory uncertainty, while that captured by the LSTM corresponds to epistemic uncertainty. While it is feasible to incorporate epistemic uncertainty for HMM, most publications employing this model overlook it. Therefore, this paper focuses on analyzing the impact of the prognostic measure on aleatory UQ.

Similarly, for LSTM, MC Dropout is often employed as a

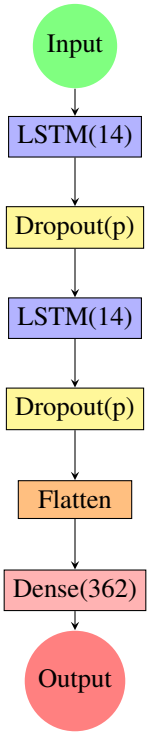


Figure 1. LSTM architecture for prognostics.

methodology to address uncertainty, hence, only epistemic uncertainty is considered. Though it is plausible to include aleatory uncertainty for LSTM, the few publications that do so, only address heteroscedastic aleatory uncertainty and disregard homoscedastic aleatory uncertainty. Consequently, this paper exclusively analyzes epistemic uncertainty via MC Dropout, a common methodology for UQ in LSTMs for prognostics.

4.3. HMM

The results for HMM both with the TI and TD prognostic measure are shown in Table 2. The results correspond to the average RMSE error and the average spread of uncertainty measured by the metric WSU, for the testing set.

Table 2. Average values of the test dataset for the prognostic performance metrics considering the TI and TD expressions of RUL for the HMM.

RUL Expression	RMSE	WSU
TI	45.00	3328839.60
TD	43.10	2978334.12

To visualize confidence intervals, engine #13 is utilized as an example. Figure 2 shows the RUL prediction alongside uncertainty quantification when employing an HMM with TI and TD prognostic measures. It is evident from the visualization that the TD approach provides results with reduced uncertainty and higher accuracy. Thus, the choice of prognostic measure significantly influences how aleatory uncertainty is quantified, as it propagates the aleatory uncertainty captured inherently by the HMM. Even with a simple model, as the HMM is, an improvement can be achieved merely by adopting a different prognostic measure. Therefore, for HMMs in prognostics, one course of action for managing uncertainty could be the development of new prognostic measures that mitigate the tendency to over-propagate inherent aleatory uncertainty captured by the HMM.

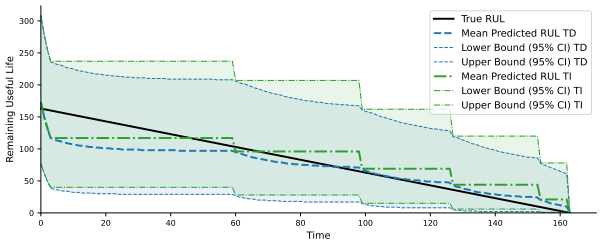


Figure 2. HMM RUL prediction for testing engine #13.

While higher performance is expected with more complex variants of HMMs, such as with a Hidden Semi Markov Model (HSMM) (Dong & He, 2007b) or the NHHSMM previously mentioned in Section 2. However, this paper offers a new time-dependent prognostic measure for the classical HMM that can be extended to other variants in future work. Furthermore, the goal of this paper is not to analyze RUL prediction accuracy but to discuss the challenges and potentials of different DDMs in terms of uncertainty.

4.4. LSTM

The results for LSTM with MC Dropout are summarized in Table 3 for dropout values 0.3, 0.6, and 0.9. The results show high accuracy in terms of RMSE for all three dropout values used, with a slightly better performance for lower dropout values. In terms of epistemic UQ, the value of WSU is higher for higher dropout values as expected.

Table 3. Average values of the test dataset for the prognostic performance metrics considering different dropout values for LSTM.

Dropout value	RMSE	WSU
0.3	1.19	150187.81
0.6	1.57	247935.42
0.9	1.63	645949.41

Figure 3 shows the RUL predictions and confidence intervals for engine #13 (the same engine used for visualization for the HMM). For clarity, only the RUL predictions with dropout values 0.3 and 0.9 are presented. The confidence intervals of the RUL predictions with the LSTM remain approximately the same throughout the degradation history since only the epistemic uncertainty is considered. Additionally, as explained in section 3, it has been claimed that MC Dropout

is not even Bayesian and the posterior distribution converges to different values depending on the dropout probability chosen by the user. In these results, it can be seen that according to the dropout values different model uncertainties are calculated. The question arises then on which is the best value to converge to the right posterior distribution of the model, meaning that there is an uncertainty on how to calculate the epistemic uncertainty.

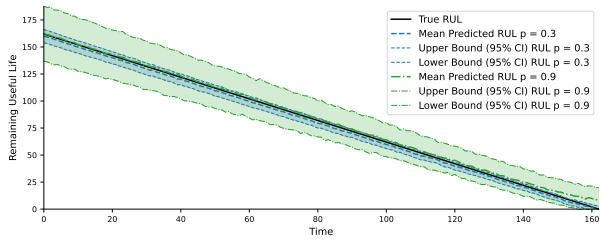


Figure 3. LSTM RUL prediction for testing engine #13 with dropout probability value of 0.3 and 0.9.

5. Future Directions in Uncertainty Quantification for Prognostics

The case study analyzed how UQ is commonly performed in prognostics for HMMs and LSTMs. While for HMMs in most cases only aleatory uncertainty is taken into account, in LSTMs only epistemic uncertainty is addressed via MC Dropout due to its simple implementation. However, even if both models could consider both aleatory and epistemic uncertainty, despite the concerns rising for both approaches, can uncertainty management be performed? Let us remember that epistemic uncertainty is reducible because it comes from the lack of knowledge. Nonetheless, it has been stated that aleatory and epistemic uncertainties often coexist, which makes it difficult to separate them (Nemani et al., 2023). By consequence, uncertainty management would not be feasible.

Hence, a different categorization of uncertainties is needed to allow differentiation. The categorization must be based on the variable of time, inherent in prognostics. This categorization should be subjective and focus on characterizing uncertainties specific to the studied system rather than uncertainties in the population. The need of a different categorization of the sources of uncertainty has been already mentioned in (Sankararaman, 2015), where the author identifies four sources of uncertainties: present, future, model and prognostic measure. The categorization was further extended in (Eleftheroglou, 2020) where a fifth source of uncertainty was included, past uncertainty.

To further explain, the five sources of uncertainty proposed are the following: first, past uncertainties are the ones that come from the manufacturing or assembly process and material quality. Second, present uncertainty refers to the lack of knowledge of the true state of health of an engineering system. Third, future uncertainty is the most difficult and important one to deal with. The future is unknown, and it is not possible to foresee the environmental conditions, loading profile, etc. Another source of uncertainty is the one from the model and it compromises several parts such as model parameters, biases, etc. The last source is the prediction method uncertainty, which is related to the uncertainty coming from the prognostic measure. In the case of supervised techniques, i.e. ML models, the model uncertainty and the prediction method uncertainty become one source.

A remark here is done for past uncertainties since they are not an uncertainty in the present, once uncertainty management is performed. For example, if sufficient data is gathered about the manufacturing process, it can be possible to manage past uncertainties and take them into account when predicting the RUL.

To the best of the author's knowledge, this categorization has not been applied to a real case study and it has only been introduced theoretically. However, an attempt to provide a better understanding on how this categorization can be implemented for HMMs is offered briefly in this section.

For HMMs, past uncertainties can be addressed by the initial parameters distributions π. Present uncertainty can be reflected by the hidden state with the highest probability at the current time step by using the forward probabilities. Future uncertainty, as already mentioned, is the most challenging one. Based on training data, loading profiles can be identified and the probability of changing from one loading profile to another one can be calculated. To account for unexpected phenomena a loading profile can be included that considers an extreme degradation rate, to give an example. Model uncertainty can be addressed by imprecise probabilities or by a sensitivity analysis, as mentioned in Section 2. Finally, the prediction method uncertainty is already considered by the prognostic measure.

DL models remain more challenging to implement under the alternative UQ categorization for prognostics due to the lack of interpretability that has already been mentioned in other publications such as (Fink et al., 2020). Given their black-box nature, the parameters of DL models do not hold a physical meaning making it intractable to connect them to each one of the five sources of uncertainties.

6. Conclusions

This paper explores uncertainty in prognostics, focusing on two main models: HMMs and LSTM networks. It finds that while HMMs primarily deal with aleatory uncertainty (inherent randomness), LSTMs predominantly address epistemic uncertainty (uncertainty from lack of knowledge).

For HMMs, results show the importance of understanding how different prognostic measures affect UQ by broaden-

ing the confidence intervals by introducing a new prognostic measurement that is time-dependant. Similarly, for LSTMs, when using the MC Dropout technique, the results show the importance of the parameter selection of the dropout probability value. Even more, from the theoretical background it has been claimed by other authors how a different dropout probability can lead to not converging to the posterior distribution needed to calculate epistemic uncertainty.

However, this paper also opens the discussion about how UQ can be used for uncertainty management in prognostics. Despite attempts to categorize uncertainty, such as distinguishing between epistemic and aleatory uncertainty, challenges persist, particularly in effectively reducing uncertainty. Future directions advocate for a different approach that considers five sources of uncertainty, such as past, present, and future uncertainties, model uncertainties, and prediction method uncertainties.

This alternative approach aims to offer a more comprehensive understanding. However, the prevalence of epistemic uncertainty poses challenges in disentangling from each one of the sources of uncertainty. Even when attempting to quantify past or model uncertainties, the presence of epistemic uncertainty persists due to data limitations and knowledge gaps. While theoretical discussions on implementing alternative categorizations for HMMs exist, practical implementation is constrained. Managing uncertainties in HMMs requires addressing multiple dimensions, encompassing past, present, and future uncertainties, as well as model and prediction method uncertainties. Conversely, implementing this alternative approach on UQ in DL models remains challenging due to their limited interpretability, raising questions about their efficacy in real-world prognostic applications.

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... Acharya, U. R. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, *76*, 243–297.

Caceres, J., Gonzalez, D., Zhou, T., & Droguett, E. L. (2021). A probabilistic bayesian recurrent neural network for remaining useful life prognostics considering epistemic and aleatory uncertainties. *Structural Control and Health Monitoring*, *28*(10), e2811.

Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, *31*(2), 105–112.

Dong, M., & He, D. (2007a). Hidden semi-markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research*, *178*(3), 858–878.

Dong, M., & He, D. (2007b). A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical systems and signal processing*, *21*(5), 2248–2266.

Eleftheroglou, N. (2020). Adaptive prognostics for remaining useful life of composite structures.

Eleftheroglou, N., Galanopoulos, G., & Loutas, T. (2024). Similarity learning hidden semi-markov model for adaptive prognostics of composite structures. *Reliability Engineering & System Safety*, *243*, 109808.

Eleftheroglou, N., Mansouri, S. S., Loutas, T., Karvelis, P., Georgoulas, G., Nikolakopoulos, G., & Zarouchas, D. (2019). Intelligent data-driven prognostic methodologies for the real-time remaining useful life until the end-of-discharge estimation of the lithium-polymer batteries of unmanned aerial vehicles with uncertainty quantification. *Applied Energy*, *254*, 113677.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Folgoc, L. L., Baltatzis, V., Desai, S., Devaraj, A., Ellis, S., Manzanera, O. E. M., ... Glocker, B. (2021). Is mc dropout bayesian? *arXiv preprint arXiv:2110.04286*.

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. arxiv 2019. *arXiv preprint arXiv:1912.02757*.

Frederick, D. K., DeCastro, J. A., & Litt, J. S. (2007). *User's guide for the commercial modular aero-propulsion system simulation (c-mapss)* (Tech. Rep.).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Giantomassi, A., Ferracuti, F., Benini, A., Ippoliti, G., Longhi, S., & Petrucci, A. (2011). Hidden markov model for health estimation and prognosis of turbofan engines. In *International design engineering technical conferences and computers and information in engineering conference* (Vol. 54808, pp. 681–689).

Le, T. T., Chatelain, F., & Bérenguer, C. (2014). Hidden markov models for diagnostics and prognostics of systems under multiple deterioration modes. In *Proceedings of the in european safety and reliability conference-esrel* (pp. 1197–1204).

Li, G., Yang, L., Lee, C.-G., Wang, X., & Rong, M. (2020). A bayesian deep learning rul framework integrating epistemic and aleatoric uncertainties. *IEEE Transactions on Industrial Electronics*, *68*(9), 8829–8841.

Nastos, C., Komninos, P., & Zarouchas, D. (2023). Non-destructive strength prediction of composite laminates utilizing deep learning and the stochastic finite element methods. *Composite Structures*, *311*, 116815.

Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A.,

... Hu, C. (2023). Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, *205*, 110796.

Pei, H., Si, X.-S., Hu, C., Li, T., He, C., & Pang, Z. (2022). Bayesian deep-learning-based prognostic model for equipment without label data related to lifetime. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *53*(1), 504–517.

Peng, W., Ye, Z.-S., & Chen, N. (2019). Bayesian deep-learning-based health prognostics toward prognostics uncertainty. *IEEE Transactions on Industrial Electronics*, *67*(3), 2283–2293.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Sankararaman, S. (2015). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing*, *52*, 228–247.

Sankararaman, S., & Goebel, K. (2013). Why is the remaining useful life prediction uncertain? In *Annual conference of the phm society* (Vol. 5).

Xiahou, T., Wang, F., Liu, Y., & Zhang, Q. (2023). Bayesian dual-input-channel lstm-based prognostics: Toward uncertainty quantification under varying future operations. *IEEE Transactions on Reliability*.

Xie, F.-Y., Hu, Y.-M., Wu, B., & Wang, Y. (2016). A generalized hidden markov model and its applications in recognition of cutting states. *International Journal of Precision Engineering and Manufacturing*, *17*, 1471–1482.

APPENDIX A

Weighted Spread of Uncertainty (WSU)

The weighted spread of uncertainty (WSU) metric is shown in 8. It calculates the area between the confidence intervals while penalizing wider confidence intervals at the end of the lifetime. The penalization is considered because the longer time that has passed, the more information is available. Variable t_i is the time unit, RUL^{upper} is the RUL value of the upper confidence interval and RUL^{lower} is the value of the lower confidence interval.

$$WSU = \sum_{i=1}^{T-1}(t_{i+1} - t_1)\left(\left(\frac{RUL_{i+1}^{upper} + RUL_i^{upper}}{2}\right) - \left(\frac{RUL_{i+1}^{lower} + RUL_i^{lower}}{2}\right)\right) \quad (8)$$

Monotonicity

The equation for the MI is provided in 9 with $y(t_i)$ as the feature value at time measurement t_i and D as the number of measurements. The results in Figure 4 show that after 20 clusters the monotonicity index converges and remains stable.

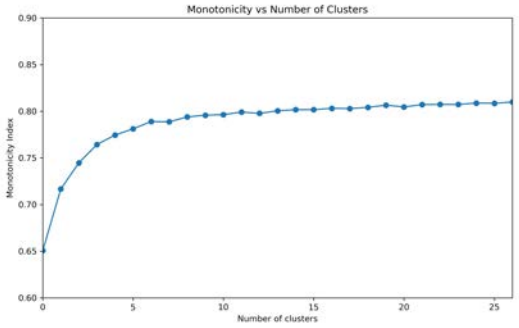


Figure 4. Monotonicity index versus the number of clusters for the sensor 11.

$$\text{MI} = \frac{\sum_i^D \sum_{j=1,j>i}^D (t_j - t_i) sgn(y(t_j) - y(t_i))}{\sum_i^D \sum_{j=1,j>i}^D (t_j - t_i)} \quad (9)$$

Bayesian Inference Criterion

In equation 10 M_i is the candidate model, $y^{(k)}$ is the sensor data from K degradation histories, $Q^{(k)}$ the state sequence for the kth degradation history, H is the number of estimated parameters of model M_i, and n the number of all the samples from the K training sessions. Figure 5 shows the results of the BIC, from which 10 states are proven to be the optimal number.

$$BIC(M_i) = \sum_{k=1}^{K} \log(P(y^{(k)}, Q^{(k)}|M_i)) - w\frac{H_i}{2}\log(n) \quad (10)$$

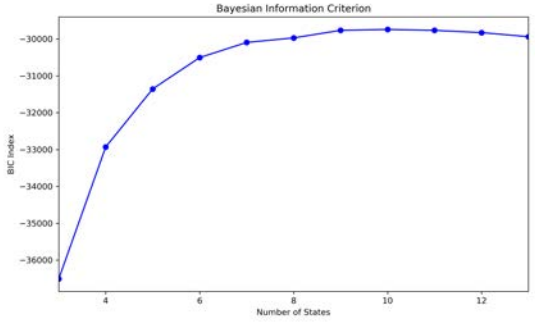


Figure 5. BIC to select the number of optimal states.

After training the HMM, the transition matrix A and the emission matrix B are estimated. The values of the elements of the matrices have been approximated

$$A = \begin{bmatrix} 0.9646 & 0.0354 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9648 & 0.0352 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9709 & 0.0291 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9735 & 0.0265 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.9591 & 0.0409 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9417 & 0.0583 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9323 & 0.0677 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9281 & 0.0719 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8941 & 0.1059 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix}
0.01 & 0.07 & 0.16 & 0.26 & 0.24 & 0.16 & 0.05 & 0.01 & 0.001 & 0.039 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0008 & 0.006 & 0.02 & 0.09 & 0.19 & 0.31 & 0.23 & 0.09 & 0.01 & 0.01 & 1.7e{-}12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.07e{-}06 & 0.003 & 0.01 & 0.05 & 0.17 & 0.29 & 0.30 & 0.11 & 0.02 & 0.003 & 0.044 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.0002 & 0.001 & 0.006 & 0.04 & 0.14 & 0.30 & 0.28 & 0.17 & 0.03 & 0.005 & 0.0278 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2.19e{-}09 & 0.003 & 0.03 & 0.16 & 0.27 & 0.31 & 0.15 & 0.05 & 0.009 & 0.018 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.004 & 0.02 & 0.10 & 0.29 & 0.31 & 0.16 & 0.06 & 0.02 & 0.36 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.002 & 0.02 & 0.07 & 0.21 & 0.27 & 0.28 & 0.09 & 0.03 & 0.028 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4.23e{-}07 & 0 & 0.005 & 0.01 & 0.10 & 0.27 & 0.30 & 0.21 & 0.07 & 0.017 & 0.018 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7.16e{-}05 & 0.005 & 0.03 & 0.12 & 0.26 & 0.28 & 0.19 & 0.07 & 0.03 & 0.012 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Unsupervised Learning for Bearing Fault Identification with Vibration Data

Gianluca Nicchiotti, Idris Cherif [1] and Sebastien Kuenlin[2]

[1]*HES-SO ▪ Haute école d'ingénierie et d'architecture, Fribourg,1700, Switzerland*
gianluca.nicchiotti@hefr.ch
idris.cherif@hefr.ch

[2]*MC Monitoring, Givisiez, 1762, Switzerland*
Sebastien.Kuenlin@mc-monitoring.com

ABSTRACT

Machine learning methods are increasingly used for rotating machinery monitoring. Usually at system set up, only data of the machinery in healthy conditions, the so-called nominal data, are available for the machine learning phase. This type of training data enables fault detection capabilities and several methods such as Gaussian Mixture Model, One Class Support Vector Machines and Auto Associative Neural Networks (Autoencoders) have been already proved successful for this task.
However, in some predictive maintenance applications, information on the type of defect may represent a key element for producing actionable information, e.g. to reduce diagnostic burden and optimize spare procurement. This requires to define classification strategies based on machine learning even in absence of data representing the behaviour of the system with defects.
In this study we present an approach that uses only nominal vibration data to train an autoencoder which will enable at same time fault identification and fault classification tasks.
As faulty data are expected to possess information content which is structured differently from the healthy ones their reconstruction at output will result inaccurate. In conventional anomaly detection approaches, the module of the reconstruction error, defined as the difference between output and input, is uses to determine an unusual input such as faults.
The proposed approach represents a step forward as here a single autoencoder is used both for detection and classification.
The underlying idea is that the components of the reconstruction error vector whose module is used to trigger fault identification in classical autoencoder approaches contain the information of the fault type. This way the analysis of the different components of the reconstruction error allows to differentiate the different types of faults.
Two methods to analyse the components of the reconstruction error vector will be discussed and their respective test results will be presented
Test data have been generated with a machine fault simulator to produce 3 different types of bearing defects with different load, speed and noise conditions. A dataset of about 10000 vibration signals has been used to evaluate the classification algorithms and to benchmark them with a supervised approach.
The results obtained using the autoencoder method do not achieve the same performances as the conventional supervised learning algorithms. However, they proved to be 88% accurate in classification when SNR is above 0dB with the ranking based method overperforming the barycentre one.

1. INTRODUCTION

Diagnostics is a crucial aspect for rotating machinery maintenance. Data processing methodologies range from traditional techniques such as frequency analysis to more innovative approaches like machine learning. In diagnostics process two main steps are often distinguished: detection and identification/classification. Detection aims to recognize the presence or absence of a defect. This can be sufficient in some situations where it is simply necessary to know if a machine is functioning correctly or if it requires intervention. Nevertheless, to optimize maintenance and repair processes, it is often essential to precisely target determining which component is failing: this requires fault identification.

Data-driven approaches are progressively more employed for anomaly detection and fault classification for machine condition monitoring purposes. However, high integrity systems could not always use the supervised learning/classification process needed for fault classification.

This depends on the fact that, at machine installation time, only healthy (nominal) data are available for training. Unsupervised learning offers a solution to fault detection by modeling nominal data and using a distance measure and a threshold for determining abnormality (Samanta, Al-Balushi, & Al-Araimi, 2003; Jack & Nandi, 2002; Booth & McDonald, 1998; Sanz, Perera, & Huerta, 2007; Guttormsson, Marks, El-Sharkawi, & Kerszenbaum, 1999; Rojas & Nandi, 2006; Prego, et al., 2013; Alguindigue & Uhrig, 1991; Fulufhelo, Tshilidzi, & Unathi, 2005; Rubio & Jáuregui, 2011). However unsupervised novelty detection approaches cannot be used for fault classification.
In (Nicchiotti et al.,2016) it has been proposed a strategy to extend machine learning capabilities from fault detection to fault classification with the constraint that only nominal data are available for training. The logic is to use a priori knowledge about the effects of each fault to be classified in order to produce input training data which are somehow fault tuned. These training data are generated by computing, on nominal data, features which are known to be the most responsive to each kind of fault which has to be classified.
The approach presented in this paper represents a step forward compared to the work presented in (Nicchiotti et al.,2016), where multiple unsupervised models were trained and classification was performed by comparing the models. In this case, classification is based on the analysis of the results of a single unsupervised model.

The case study used to validate this new approach is the classification of faults in ball bearings with a machine learning approach where only healthy data are used for training. The study required taking measurements on defective bearings under various operating and noise conditions. The collected signals were preprocessed to extract training features both in time (RMS, Kurtosis, Crest Factor, etc.) and frequency domains. The frequency domain proved to be particularly effective in discriminating between different types of failures, due to the characteristic frequencies associated with the defects.

The paper is organized as follows. Next section will briefly illustrate the test rig and the data set characteristics.
A description of data-driven method used in this study will be presented in section 2. The focus will be on Auto Associative Neural Networks (AANN) .
The novel strategy to extend the data-driven capabilities from detection to classification will be described in section 4
Two methods for classifying defects will be explored: the first based on the ranking of reconstruction errors components, the second on the analysis of the barycenter of the reconstruction error when represented in a polar plot.
The results obtained with the 2 data-driven methodologies will be then compared and discussed and their robustness against noise characterized.
The classification results will be finally benchmarked against supervised approaches.

2. Acquisition Setup And Dataset

The signals were acquired using acquisition systems developed by MC-Monitoring. Measurements were conducted on a fault simulator, allowing for measurements under different operating conditions. The fault simulator enables the rotation of bearings under various load, unbalance, and speed conditions. Bearings can be affected by defects in the inner race, outer race, balls, and a case presenting a combination of defects. The defective bearings were placed at location 5 (see figure 1), and the measurement via the accelerometer is carried out along the x and y axes. The sampling rate was 50 kHz

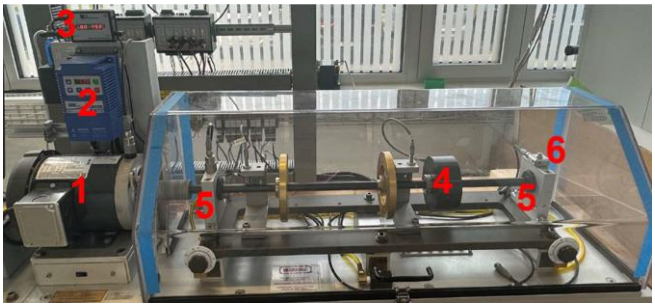


Figure 1. Machine Fault Simulator and acquisition setup. 1 AC Motor, 2 Frequency Converter ,3 Tachometer, 4 Additional mass load, 5 Right and left bearing 6Acceleration sensor x-y, 100mV/g.

For each type of bearing fault and under different loads and speed conditions, a 6-minute acquisition of the vibration signal has been performed.

After digital conversion, the raw signal undergoes filtering to remove the DC component. We then calculate the Root Mean Square (RMS) value of the filtered signal, which becomes the reference point for adding white noise. To assess the process's robustness, seven levels of white noise were added to each original signal. This resulted in a total of 140 sequences, each containing 6 minutes of healthy and faulty signal data under different load, speed, and signal-to-noise ratio (SNR) conditions.

3. Machine Learning Methods

Machine learning offers diverse tools for monitoring machine health, including density methods (KNN), boundary methods (SVM), and reconstruction methods (AANN) (Johannes, 2001). These techniques have been successfully applied to fault detection (Samanta et al., 2003; Jack & Nandi, 2002; Booth & McDonald, 1998). When used for classification, these approaches all require pre-existing fault data for training (Alguindigue & Uhrig, 1991; Fulufhelo et al., 2005; Wang et al., 2020; Prego et al., 2013).

However, acquiring such data poses a challenge when dealing with new equipment. In such scenarios, data-driven anomaly detection emerges as the only possible alternative which is

exemplified in studies by Rubio & Jáuregui (2011), Guttormsson et al. (1999), and Sanz et al. (2007) . Methods like Auto-Associative Neural Networks (Sanz et al., 2007) and one class SVM (Guttormsson et al., 1999) are among the most widely used methodologies that rely on "one-class classification", when only healthy data is available for training.

Despite their success in fault detection, one-class classification methods haven't been explored for fault identification. This research aims to bridge that gap by incorporating expert knowledge ("a priori") into these data-driven ("a posteriori") techniques, implementing fault classification within the AANN framework.

3.1. AANN

Auto-Associative Neural Networks (AANNs), also known as Replicator Neural Networks or Autoencoders, are like smart copycats in the world of artificial intelligence. These networks are trained to mimic whatever data they're given, but with a twist: they have a hidden layer with fewer neurons than the input and output. This "bottleneck", shown in Figure 1, forces them to compress the information, essentially learning the key features of the data they're trained on.

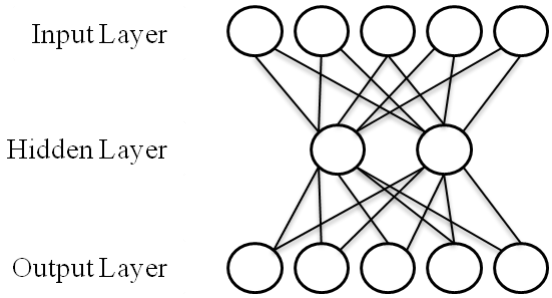


Figure 2. AANN Architecture.

Imagine training an AANN with healthy equipment data. Once trained, it can accurately reproduce similar healthy data. However, faulty data will contain different patterns that the AANN struggles to compress in the bottleneck. This results in a larger reconstruction error, which is the difference between the original data and the AANN's attempt to recreate it.

The reconstruction error can be considered as a measure of strangeness. The higher the error, the more different the data is from what the AANN knows as "healthy." By setting a threshold for this error, we can create a simple fault detection system: anything with an error above the threshold is likely faulty.

In practice once a new sample is processed by the AANN, the measure of the difference between output and input vectors, the Reconstruction Error (RE) of an input vector X, is computed as

$$RE = \|X - O_X\| \quad (1)$$

where O_X is the output of the AANN and || symbol stands for any p-norm. Once computed the R_E, a fault or anomaly detection logic can be easily implemented for instance by thresholding.

4. FAULT IDENTIFICATION STRATEGY

This section aims to examine the usage the reconstruction error of an autoencoder to classify different types of defects.

To compensate for the lack signals associated with the defects, the idea is to leverage the "a priori" knowledge of the phenomena linked to the type of fault and encode it within the autoencoder process. As shown in figure 3, this process requires an initial step consisting of extracting features from the signal through appropriate signal processing techniques.

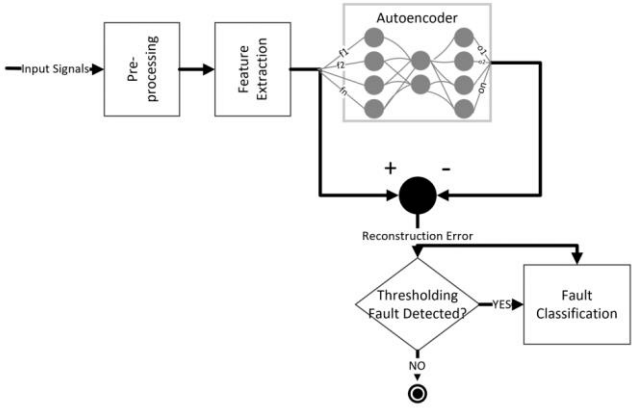


Figure 3. Main Processing Steps.

4.1. Pre-processing

In the subsequent discussion, the term "feature" refers to an individual measurable attribute of the observed phenomenon. In this research, the features represent various characteristics of the signals extractable from the signals. Features represent unique clues about a machine's condition, like symptoms for a doctor. Think of these features as different dials on a dashboard, each showing a different aspect of the machine's health.

Raw sensor data from ball bearings needs preparation before computing features and feeding it into machine learning algorithms. This involves filtering, windowing, and extracting the signal's "envelope".

- Filtering - Since the information we care about lies above 2 kHz, lower frequencies are irrelevant and clutter the analysis. We utilize a Butterworth bandpass filter [2-22kHz]to selectively remove them. In this proof of concept, we decided to use such a large bandwidth to represent the worst case. Practically, it is more efficient to filter around the frequency resonance of the entire system (motor, bearing, sensors, etc..) which is between generally somewhere between 2kHz and 20kHz to minimize the noise. However, as we wanted a "generic" system, we decided to use the overall bandwidth of our acquisition system. This also presents a practical benefit of not requiring to configure the filter during the installation procedure.

- Envelope Extraction - Ball bearing vibrations, like the one shown in Figure 3, contain information in their "envelope". This envelope reflects the modulation of the bearing's natural resonance frequency caused by impacts between rolling elements and defects. To uncover the characteristic frequencies of these defects, we calculate the Fast Fourier Transform (FFT) on the extracted envelope, not the raw signal. As most of the rotating machines we monitor in our applications run between 25Hz and 60Hz, we know that the characteristic frequency of the faults (BPFO, BPFI, etc) are between, let say, 5Hz an 500Hz, so the envelope size was chosen according to these values.

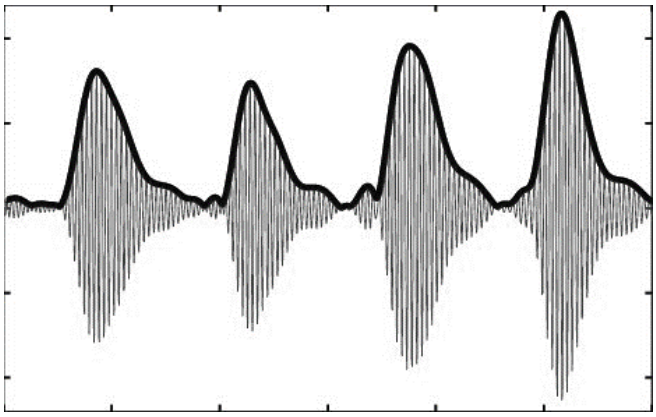


Figure 4. Bearing signal: Raw (slim lines) Envelope (bold lines)

- Windowing - To enhance the accuracy of frequency analysis, we apply a Hann window to the filtered data. This window smooths the signal edges, reducing artifacts in the resulting spectrum. The measured signal (240 s) will be divided into 1-second samples with a 1/2-second overlap. The choice of 1 second allows balancing the need for enough signals for machine learning model training and retaining sufficient characteristic information of the vibrational signal generated by the bearing. A too small window would make the model unreliable if the pseudo-periodic nature of the signal is not preserved in the sample. A too large window would also not allow good model generalization due to the more limited training dataset.

4.2. Feature extraction

Features have been extracted with time domain and frequency domain analysis.

In the time domain, signals from defective bearings exhibit periodic impulses corresponding to impacts between the balls and the cage defect or between the defective ball and the metallic components. The impulses excite the resonance frequencies of the system. Each impact can be compared to the impulse response of the system due to the short duration of contact between a ball and the defect The presence of defects in a machine can be detected by analysing the vibration signal. Defects increase the energy of the vibration signal and modify its statistical distribution. These changes can be used to identify the presence and severity of the defects. Time domain features used in this study are (Hornavar & Martin, 1995): RMS,Crest Factor, Kurtosis, Skewness, Impulse Factor and Form Factor.

Each bearing has a unique "fingerprint", its characteristic frequencies, determined by its geometry and rotation speed. (Kamaras et al. ,1995, Andhare, 2010)

- FTF - Fundamental Train Frequency: This is the rotation frequency of the bearing cage.

- BPFI - Ball Pass Frequency of the Inner Race: This frequency is generated by the passage of balls over the inner ring.

- BPFO - Ball Pass Frequency of the Outer Race: This frequency is generated by the passage of balls over the outer ring.

- BSF - Ball Spin Frequency: This frequency is related to the rotation of the balls.

When a fault develops, these specific frequencies become amplified, acting like warning lights. These frequencies not only reveal the presence of a problem but also pinpoint the exact type of fault, allowing for targeted repairs and preventing unnecessary downtime. Hence to identify faults in bearings, features were extracted from the vibration signal's envelope spectrum. These features included the peak ratio amplitudes of characteristic frequencies related to bearing health: the Ball Pass Frequency Outer Race (BPFO), Ball Pass Frequency Inner Race (BPFI), and Ball Spin Frequency (BSF). Additionally, the spectral centroid and the energy in the band 10-20 kHz were included. These features formed the frequency "fingerprint" of the bearing's health and were fed into an Auto Associative Neural Network (AANN) for fault classification.

To extract the features, we first chopped the time signal into half second intervals, making sure to overlap them by 0.25s to capture any important transitions. From each chunk, we

then extracted 14 specific features which we fed into the autoencoder for further analysis.

4.3. Classification

The fault detection by autoencoder is based on the premise that the reconstruction error for data similar to those used for training will be lower than the error for data from faulty bearings.

Since different features do not all have the same range of values, the data has to be standardized (Equation 2). Each new set of tested data will be standardized using the mean and standard deviation of the training data.

$$F_i = \frac{f_i - \mu_i}{\sigma_i} \qquad (2)$$

Where f_i is the value to be standardized, μ_i , σ_i are the mean and the standard deviation of the training set for feature i and F_i is the standardized feature value.

The parameters of the autoencoder have been determined according to the average reconstruction error on healthy data. The autoencoder has a single hidden layer which contain 10 neurons This value represents a good compromise between low reconstruction error and moderate training time. The maximum number of iterations (Epochs) for training is set to 500 beyond this, the improvement in performance is not significant.

For each feature F_i at the input, the autoencoder calculates a corresponding output O_i . The difference $E_i = \|F_i - O_i\|$ represents the components of the reconstruction error along the various axes represented by the features used, as shown in Figure 5.

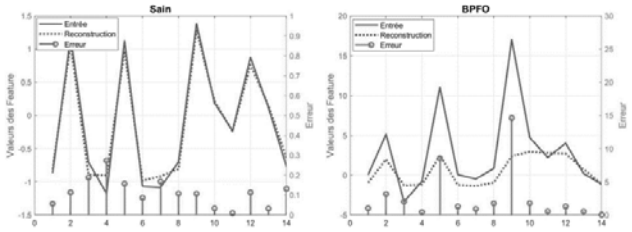


Figure 5: Reconstruction error for a healthy signal (left) and outer race fault (right). X-axis represent the feature index Errors E_i in stems (vertical lines), continuous lines input features dotted features as reconstructed by autoencoder

The two classification strategies here presented assume that the relative values of components E_i of the reconstruction error depend on the type of defect.

The first approach maps E_i on a polar plot (Figure 6), and uses the angle of the reconstruction error to discriminate the different types of faults.

To differentiate between fault types, each feature i is assigned a specific angle (θ_i). Using a priori knowledge about fault behavior, these angles are carefully chosen to maximize the angular separation between features clearly associated with different faults (e.g., Inner Race, Outer Race, Ball). For example, BPFO, BPFI, and BSF might be assigned 0°, 120°, and 240° directions, respectively.

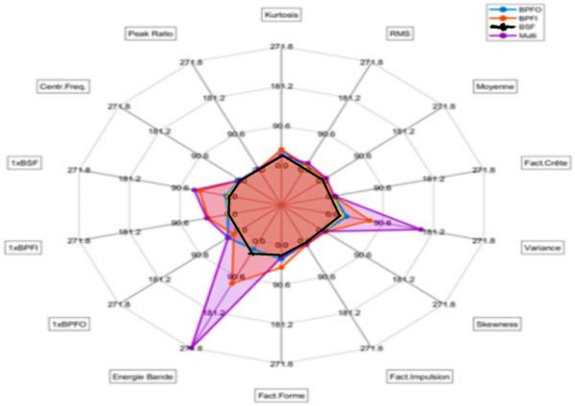


Figure 6: Polar Plot of the reconstruction error. Blue contour represents the shape of outer race fault, Red of inner race, Black ball fault, Violet Multiple faults. Different faults types correspond to different contours.

Next, each feature's reconstruction error component E_i is represented as a 2D vector $\overrightarrow{E_i}$ where the magnitude reflects the error value of the component itself and the angle corresponds to the pre-assigned θ_i based on fault type a priori knowledge. By summing these vectors, a resulting reconstruction vector $\vec{V}$ is created

$$\vec{V} = \sum_{i=1}^{N} \overrightarrow{E_i} \qquad (3)$$

Based on the vector direction Θ of $\vec{V}$ *with* Θ=arg($\vec{V}$) the fault type can be classified.

For instance, if Θ is falling within -60° to 60° we can classify the fault as inner race between 60° and 180° as outer race, else as ball error. The underlying idea is that reconstruction errors not strictly dependent on the type of defect interfere destructively, highlighting the direction of the defect in the polar plot.

The second method assumes different fault types make it harder for the autoencoder to reconstruct certain feature components. By leveraging a priori knowledge, we anticipate the ranking of reconstruction errors magnitude for the features across different fault scenarios as shown in Table 1.

Table 1. Ranking fault signatures

Fault type	Reconstruction error $E_i = \|F_i - O_i\|$													
	←Bigger													Smaller →
Inner	1	8	10	5	9	12	7	2	11	6	3	4	13	14
Outer	9	5	12	8	11	1	10	2	7	6	3	4	13	14
Ball	12	8	1	5	9	7	6	10	11	2	3	4	13	14

This expected ranking serves as a signature to identify the actual fault based on the observed ranking of reconstruction errors. Comparing the actual observed order to the expected ranking enables the classification task.

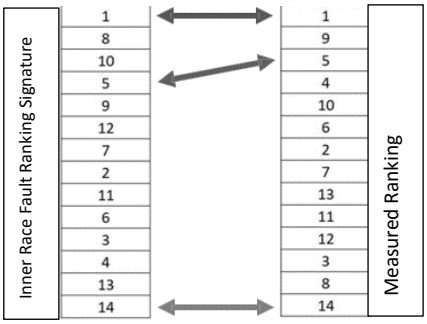


Figure 7: Example of ranking signature for Outer Race Fault and RBO algorithm

To effectively compare ranking list and classify the fault type, the algorithm needs to consider the varying importance of features in their ranked order (see Figure 7). The correspondence between two top-ranked features (top arrow) is more important than two lower-ranked features (bottom arrow). Additionally, two identical features that are not at the same rank (center arrow) must be taken into account. To accomplish this, the Rank Biased Overlap (RBO) algorithm (Joshi 2021), which meets these requirements, was used to compare the rankings. This algorithm allows assigning a weight (p) more or less significant to elements at the top of the ranking. The result of the comparison between the two lists is a number between 0 and 1 (the value '1' is obtained for two identical lists).

5. TEST RESULTS

An autoencoder with 14 input features and a single hidden layer with 8 nodes to distinguish healthy and faulty system states has been trained with 3000 healthy samples, each represented by a vector of 14 features extracted from a 0.25-second signal window. The model's performance was evaluated on 3000 healthy and 9000 faulty samples. While confusion matrices provided insights into classification errors, this document focuses on precision, defined as the ratio of the total number of correct predictions to the total number of predictions made by the model.

Figure 8 demonstrates the model's ability to accurately detect healthy states with excellent precision even when dealing with high levels of noise in the signal.

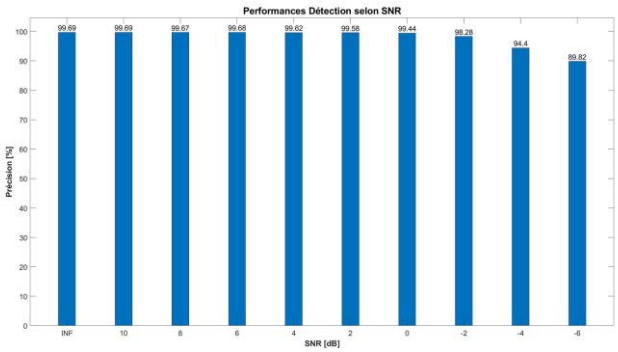


Figure 8: Precision of the anomaly detection as a function of SNR(dB)

To detecting faults, we set a threshold (Th_D) based on the reconstruction error (RE). If RE exceeds Th_D, a fault is likely present. This threshold is calculated as the mean (μ_{TS}) plus three times the standard deviation (σ_{TS}) of the reconstruction error computed on the training data (Equation 4).

$$Th_D = \mu_{TS} + 3 \cdot \sigma_{TS} \quad (4)$$

The autoencoder demonstrates exceptional anomaly detection capabilities even under low noise conditions, achieving precision levels exceeding 99.4% for signal-to-noise ratios (SNR) up to 0 dB. However, as noise levels increase, performance drops and false alarms become a concern, at SNR 6 dB, precision falls below 90%.

5.1. Vector Direction classification

Initially, with 14 features, the method based on the direction of the reconstruction vector (Θ=arg($\vec{V}$)) only achieved a 76.8% precision. The unsatisfactory percentage is likely due to the correlation between some features which produced the same interference pattern. Therefore, we applied Principal Component Analysis (PCA) (Shlens, 2014) to identify redundancies and reduce the number of features to 8 This process ends up with the selection of the most informative features, in statistical sense, like RMS, Kurtosis, Peak Factor, Impulse Factor, and key frequency components). This dimensionality reduction significantly improved classification performance, boosting the precision to 87.94%.

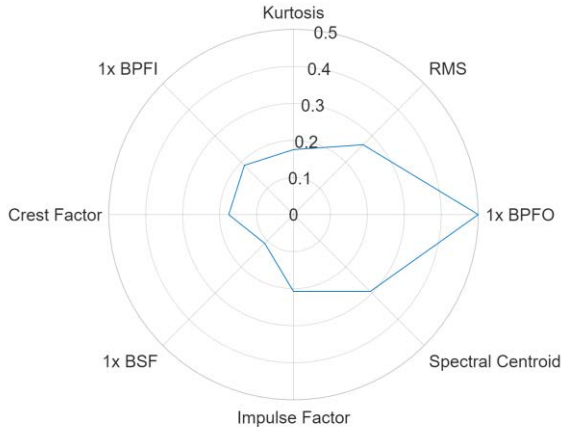
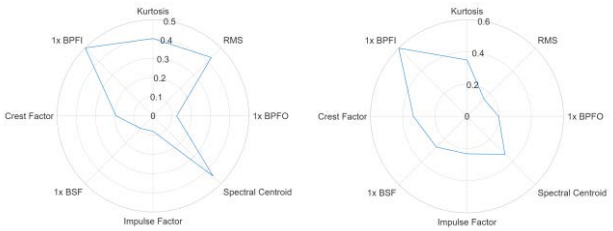


Figure 9: Polar Plot with 8 features pour outer race fault



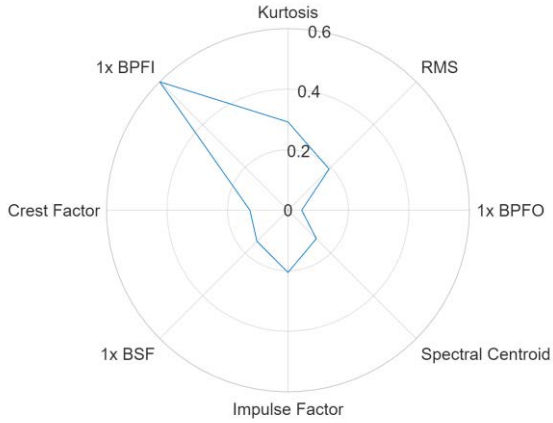


Figure 10: Polar Plots with 8 features for an inner race fault

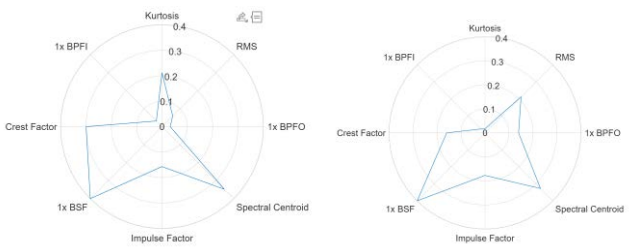


Figure 11: Polar Plots with 8 features for a ball fault

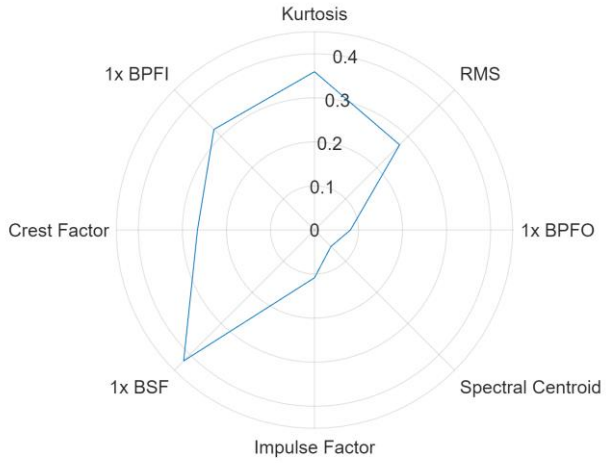


Figure 12: Polar Plots with 8 features for a ball fault, with misclassification

In addition to automated fault classification, our method provides a visual tool for identifying different fault types. Figures 9-11 showcase polar diagrams where each plot displays examples correctly classified. Figure 12, however, depicts a scenario where misclassification occurred.

In some maintenance situations, where misidentifying a failure could have serious consequences, operators can leverage these visualizations to perform a semi-automated diagnosis, especially if the automated results lack sufficient confidence. This allows them to combine the model's insights with their own expertise for a more informed decision.

5.2. Rank Order classification

When using the rank order RBO method for classification, the initial precision with 14 features was only 41.46%. Similar to the previous approach, we reduced the number of features to decrease redundancy and eliminate less relevant information. Using the same features as the "vector direction" classifier, the RBO method improved, and achieved a maximum precision of 76.8% when the parameter p (Joshi 2021) is set to p =0.9. The parameter p determines the weighting of the first positions in the similarity measurement between two ranked lists. By adjusting the value of p, it is possible to control the importance given to the first positions compared to the subsequent positions, thus providing flexibility to evaluate the similarity between ranked lists according to different criteria. The value of p is chosen between 0 and 1. In this case, it has been decided to give more importance to the first positions in the ranking, which have a more significant impact on the classification performances.

Therefore, the « directional » classifier appears more effective in our tests.

While unsupervised methods like RBO offer an advantage in not requiring labelled data, their precision in this case

(76.8%) falls short compared to supervised learning (>99% precision). (Cherif, 2023) This is likely due to the inclusion of actual fault data in the training set for supervised methods, providing them with a more direct understanding of failure patterns.

6. CONCLUSIONS

The application of unsupervised machine learning techniques for detecting bearing faults and anomalies in rotating machinery offers a compelling array of advantages for practical business implementations. By eliminating the need for labeled faulty data, these approaches streamline the deployment process, making it more accessible and cost-effective for industrial settings.

The ability to operate without pre-existing fault data enables unsupervised algorithms to uncover previously unrecognized patterns and anomalies, providing early detection of faults and proactive maintenance opportunities. This early detection capability, coupled with scalability and adaptability to changing conditions, empowers businesses to enhance reliability, minimize downtime, and optimize maintenance strategies.

In essence, leveraging unsupervised machine learning in industrial contexts not only circumvents the challenges associated with acquiring labeled data but also delivers tangible benefits in terms of reliability, efficiency, and cost-effectiveness. This approach represents a transformative paradigm for bearing fault detection and anomaly monitoring, enabling businesses to proactively manage their assets and maximize operational performance without relying on historical fault data.

However, whilst supervised and unsupervised methods show similar performance in defect detection, our proposed approach using an autoencoder for classification falls short compared to supervised learning. However, our method offers the crucial advantage of not needing rare defect data for training. This combination of unsupervised anomaly detection and classification enables defect detection without labeled data.

The "directional error" method achieves a promising 87.94% accuracy through optimized feature selection.

To further improve our classification system, we are pursuing two complementary research directions:

1. Improving precision over time: We're exploring how to combine classification results over time to potentially boost precision.
2. Implementing a rejection logic (Bartlett & WegKamp, 2008; Chow ,1970): This framework aims to prevent critical misclassifications by allowing the model to avoid predictions when uncertain, at a predefined cost. This enables semi-automatic diagnostics where polar plots of ambiguous cases can be sent to operators for confirmation.

From the validation point of view, it has been planned to validate our methodology on HUST bearing dataset (https://data.mendeley.com/datasets/cbv7jyx4p9/3)

NOMENCLATURE

E_i	Reconstruction error component
Θ	Direction of Reconstruction Vector
p	Weight of RBO
RE	Reconstruction Error
Th_D	Detection Threshold
$\vec{V}$	Reconstruction vector

REFERENCES

Alguindigue, I., & Uhrig, R. E. (1991). Vibration monitoring with artificial neural networks. *Tennessee: OECD publishing.*

Andhare, A. (2010). Condition Monitoring of Rolling Element Bearings. *Lap Lambert Academic Publishing GmbH KG*. p.9-11, 65-72.

Bartlett P. L. and Wegkamp M. H. (2008) Classification with a reject option using a hinge loss. *Journal of Machine Learning Research, 9:*1823–1840, 2008.

Bilmes, J. A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *Berkeley: International computer science institute.*

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *New York: Springer Science+Business Media.*

Booth, C., & McDonald, J. R. (1998). The use of artificial neural networks for condition monitoring of electrical power transformers. *Neurocomputing, 97-109.*

Cherif I. (2023) Détection des défauts de roulements par analyse de vibrations. *Travail de Bachelor Haute école d'ingénierie et d'architecture, Fribourg*

Chow. C. K. (1970) On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Fulufhelo, V. N., Tshilidzi, M., & Unathi, M. (2005). Early classifications of bearing faults using hidden Markov models, Gaussian mixture models, Mel- frequency cepstral coefficients and fractals. *International Journal of Innovative Computing, Information and Control*, 1281-1299.

Guttormsson, S., Marks, R., El-Sharkawi, M., & Kerszenbaum, I. (1999). Elliptical novelty grouping for on-line short-turn detection of excited running rotors. *Energy Conversion, IEEE Transactions on*, 16-22.

Honarvar, F. and Martin, H.R. (1995) Application of statistical moments to bearing fault detection. *Applied acoustics*, 44 : p.67-78,

Jack, L. B., & Nandi, A. K. (2002). Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms. *Mechanical Systems and Signal Processing*, 373-390.

Johannes, M. D. (2001). One-class classification. *Delft: Advanced School for Computing and Imaging*.

Joshi, P. "RBO v/s Kendall Tau to compare ranked lists of items". Towards Data Science. *Retrieved 23-01-2024.: https://towardsdatascience.com/rbo-v-s-kendall-tau-to- compare-ranked-lists-of-items-8776c5182899*

Kramer, M. A. (1992). Autoassociative neural networks. *Computers & chemical engineering*, 16(4), 313-328.

Kamaras, K., Garantziotis, A., & Dimitrakopoulos, I. Vibration Analysis of Rolling Element Bearings (Air Conditioning Motor Case Study). *Retrieved January 25, 2023 https: //fnt.com.cy/ images/ Rolling%20Element%20Bearings%20Vibration %20Analysis.pdf*

Nicchiotti, G., Fromaigeat, L., & Etienne, L. (2016). "Machine Learning Strategy for Fault Classification Using Only Nominal Data". *PHME Conference 2016*

Ng, A. (2015, September 21). Machine Learning Course Materials. *Retrieved January 15, 2023, from http://cs229.stanford.edu/materials.html*

Prego, T. d., de Lima, A. A., Netto, S. L., da Silva, E.A., Gutierrez, R. H., Monteiro, U. A.,Vaz, L. (2013). On Fault Classification in Rotating Machines using Fourier Domain Features and Neural Networks. *Circuits and Systems (LASCAS), 2013 IEEE Fourth Latin American Symposium n*, 1-4.

Rojas, A., & Nandi, A. K. (2006). Practical scheme for fast detection and classification of rolling-element bearing faults using support vector machines. *Mechanical Systems and Signal Processing 20*, 1523-1536.

Rubio, E., & Jáuregui, J. C. (2011). Time-Frequency Analysis for Rotor-Rubbing Diagnosis. In F. Ebrahimi, *Advances in Vibration Analysis Research* (pp. ISBN: 978-953-307-209-8, InTech, DOI:10.5772/15186).

Samanta, B., Al-Balushi, K. R., & Al-Araimi, S. A. (2003). Artificial neural networks and support vector machines with genetix algorithm for bearing fault detection. *Engineering Applications of Artificial Intelligence*, 657-665.

Sanz, J., Perera, R., & Huerta, C. (2007). Fault diagnosis of rotating machinery based on auto-associative neural networks and wavelet transforms. *Journal of Sound and Vibration 302*, 981-999.

Shlens, J. (2014). A tutorial on principal component analysis. Retrieved 14.2.2024 *https://doi.org/10.48550/arXiv.1404.1100*

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., & Smola, A. J. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation, 13*(7), 1443-1471.

BIOGRAPHIES

Gianluca Nicchiotti received an MSc in Physics from Università di Genova, Italy in 1987 and an MSc in Applied Science from Cranfield University UK in 2013. Since 2022 he holds the chair of Signal Processing at Haute école d'ingénierie et d'architecture in Fribourg. From 2005 to 2022 worked in the aerospace domain as system engineer by Meggitt, Switzerland. Prior to Meggitt, he managed an image processing research team at Elba Research Center and developed algorithms for sport video special effects at Dartfish. His career started at Elsag Bailey R&D in underwater acoustic cameras field. He is author of more than 50 papers and 5 patents. His current research interests are quantum computing and machine learning.

Idris Cherif got a degree in Electronic Engineering from Haute école d'ingénierie et d'architecture in Fribourg now he is R&D engineer at Inergio.

Sebastian Kuenlin's received a degree in Electronic Engineering from Haute école d'ingénierie et d'architecture in Fribourg, Switzerland, where he earned a Bachelor of Science HES in Electrical Engineering with a specialization in Electronics in 2010. At the end of his studies, he conducted his diploma work at Lawrence Berkeley National Laboratory, focusing on testing stave/module systems for the ATLAS detector at CERN. During this time, he developed an electronic test bench based on FPGA. Transitioning to MC-monitoring SA in 2011, Sébastien worked in embedded system development before assuming leadership roles, ultimately becoming Chief Technology Officer and of MC-monitoring SA in 2018. His current areas of expertise are acquisition system and signal processing.

Virtual Sensor for Real-Time Bearing Load Prediction Using Heterogeneous Temporal Graph Neural Networks

Mengjie Zhao[1], Cees Taal[2], Stephan Baggerohr [3], and Olga Fink[4]

[1,4] *Intelligent Maintenance and Operations Systems, EPFL, Lausanne, Switzerland*
mengjie.zhao@epfl.ch
olga.fink@epfl.ch

[2, 3] *SKF, Research and Technology Development, Houten, the Netherlands*
cees.taal@skf.com
stephan.baggerohr@skf.com

ABSTRACT

Accurate bearing load monitoring is essential for their Prognostics and Health Management (PHM), enabling damage assessment, wear prediction, and proactive maintenance. While bearing sensors are typically placed on the bearing housing, direct load monitoring requires sensors inside the bearing itself. Recently introduced sensor rollers enable direct bearing load monitoring but are constrained by their battery life. Data-driven virtual sensors can learn from sensor roller data collected during a battery's lifetime to map operating conditions to bearing loads. Although spatially distributed bearing sensors offer insights into load distribution (e.g., correlating temperature with load), traditional machine learning algorithms struggle to fully exploit these spatial-temporal dependencies. To address this gap, we introduce a graph-based virtual sensor that leverages Graph Neural Networks (GNNs) to analyze spatial-temporal dependencies among sensor signals, mapping existing measurements (temperature, vibration) to bearing loads. Since temperature and vibration signals exhibit vastly different dynamics, we propose Heterogeneous Temporal Graph Neural Networks (HTGNN), which explicitly models these signal types and their interactions for effective load prediction. Our results demonstrate that HTGNN outperforms Convolutional Neural Networks (CNNs), which struggle to capture both spatial and heterogeneous signal characteristics. These findings highlight the importance of capturing the complex spatial interactions between temperature, vibration, and load.

1. INTRODUCTION

Bearings are essential components in mechanical systems, ensuring the efficient and reliable operation of machinery and equipment across diverse industries, including wind energy, aerospace, and automotive sectors. Real-time monitoring of bearing conditions is crucial for optimal performance and proactive maintenance (Hou & Wang, 2021). Knowing the actual load experienced by bearings offers several key benefits. Firstly, deviations from the original design loads signal the need for adjustments to operational parameters and maintenance schedules. This allows for proactive prescription of health-aware load profiles, potentially extending the bearing's service life. Moreover, load monitoring aids in early detection of misalignments, enabling timely proactive adjustments to prevent further damage (Widner & Littmann, 1976). Additionally, knowledge of bearing load facilitates more accurate diagnosis of potential bearing faults (Peng et al., 2020). Finally, bearing loads are a key factor influencing bearing lifespan and failure (Harris & Kotzalas, 2006), and their understanding enables predicting damage propagation (Morales et al., 2019). An in-depth understanding of the load is essential for accurate Remaining Useful Life (RUL) prediction and effective Prognostic and Health Management (PHM).

Directly measuring bearing loads during operation presents complex challenges. Traditional approaches, typically using strain gauges, require direct contact or close proximity to the bearing's rolling elements. This introduces significant logistical and technical hurdles (Konopka et al., 2023), including accessing power and establishing sensor communication, making installation more expensive than conventional condition monitoring sensors, such as for vibration and temperature.

Recently, wireless sensor roller technology has been introduced, wherein sensors are embedded inside a rolling element to allow in-operation measurement of bearing loads (Baggerohr,

2023). However, their utility is still constrained by battery life. A virtual sensor could overcome this limitation by providing continuous, long-term load predictions, even when the sensor roller's battery is depleted. Specifically, our goal is to develop a virtual sensor that maps the measurements of conventional bearing condition monitoring sensors to loads. Since the relationships between these sensors and load are influenced by factors such as stiffness, damping, and thermal behavior, and are often unknown in real-world applications, we adopt a data-driven approach. Sensor roller provides crucial ground-truth load data, which is significant for enabling the development of this virtual sensor. Estimating the load without such direct data is extremely difficult without extensive modeling. Our approach not only extends the value of the physical sensor roller but also supports advanced PHM.

Virtual sensors have been applied in many different applications ranging from environmental sensing to complex industrial systems. They leverage readily available measurements and computational models to infer quantities that are challenging or costly to measure directly (Martin et al., 2021). They also play a crucial role in digital twins, providing insights beyond what physical sensors can capture (Song et al., 2023). Two primary directions exist for virtual sensors: *model-based* and *data-driven*. Model-based approaches rely on well-defined physical laws and principles to develop models describing the system of interest. In contrast, *data-driven* approaches use machine learning and data mining algorithms to find patterns and relationships within sensor data. Model-based virtual sensors require using existing sensor data to accurately infer and update model parameters to ensure accurate estimations. Methods such as Kalman filtering, which dynamically updates model states in real-time based on noisy sensor measurements, are well-established for calibrating physics-based virtual sensors for load estimation models (Kerst et al., 2019). Alternatively, Gaussian processes can be applied to latent force models to infer unknown load dynamics from a sensor network (Bilbao et al., 2022). While powerful, these methods rely on prior knowledge of the system's physics, which can be challenging or infeasible to obtain in many real-world cases. In contrast, data-driven virtual sensing offers flexibility by directly learning complex relationships from data. For example, (Dimitrov & Göçmen, 2022) demonstrated the potential of Long Short-Term Memory (LSTM) networks for predicting wind turbine blade root bending moment using SCADA data. (Wang et al., 2021) developed a Deep Belief Network (DBN) with event-triggered learning (DBN-EL) to improve the efficiency and accuracy of a water quality soft-sensing model for the wastewater treatment processes from the sensor data.

Model-based methods often depend on prior knowledge of the system's physics. In contrast, data-driven approaches can overcome this limitation but may require other forms of ground truth to learn the functional relationships, such as simulation data (Dimitrov & Göçmen, 2022) or periodic lab-based measurements (Wang et al., 2021), making them difficult to apply in real-world scenarios. Fortunately, bearing sensor rollers allow direct measurement of bearing load in operation, offering a direct ground truth that enables us to learn the complex relationships between load and conventional bearing condition monitoring sensors through supervised learning.

For large-size bearings, such as main shaft bearings in wind turbines, a common approach for bearing condition monitoring involves positioning multiple sensors around the bearing to measure rotational speed, vibration, and temperature. Although a correlation exists between load and these sensor readings, the relationships are complex and difficult to model accurately due to the lack of exact physical models. However, there exists an additional inductive bias in the form of spatial information, such as the correlation between higher temperatures and areas of increased load. Leveraging this spatial information can offer valuable insights into load distribution. While traditional machine learning algorithms struggle to effectively utilize this spatial information, Graph Neural Networks (GNNs) are well-suited for handling spatial-temporal dependencies (Jin et al., 2023). By modeling sensors and their connections as a graph, GNNs can directly capture the spatial dependencies and relationships between different sensor readings. They utilize message-passing techniques, where information from neighboring sensors is iteratively processed and aggregated, building a global understanding from local information (Gilmer et al., 2017). GNNs have been successfully applied in areas such as bearing remaining useful life prediction (Yang et al., 2022), cyber-physical attack detection for water distribution systems (Deng & Hooi, 2021), sensor calibration for air pollution (Niresi et al., 2023) and fault detection for chemical process plants (Zhao & Fink, 2023).

Nevertheless, existing GNN methods often assume relatively similar feature characteristics across nodes. Although GNNs have been applied to heterogeneous sensor networks, the focus has typically been on handling different sensor types (e.g., temperature, humidity, pressure). In these cases, while the data originates from diverse sources, the signal characteristics often exhibit some similarities. In our scenario, the heterogeneity is in signal characteristics. Vibration and temperature signals exhibit very different dynamics and frequencies. This poses a novel and significant challenge for GNNs, which often struggle to effectively integrate and learn from such highly diverse signal characteristics.

To address the challenge of heterogeneous sensor characteristics, we propose a novel virtual load sensor based on Heterogeneous Temporal Graph Neural Networks (HTGNNs). By explicitly modeling high and low-frequency signals as distinct node types and differentiating their interaction types, our HTGNN effectively fuses the information from diverse sensors. This enables more accurate load prediction, overcoming

the limitations of traditional GNNs. To the best of our knowledge, this represents the first design of such an architecture to analyze diverse sensor types for bearing load estimation.

The remainder of this paper is organized as follows: Sec. 2 describes the task of a bearing virtual sensor. Sec. 3 elaborates on HTGNN's core components to model the heterogeneous dynamic relationships within the bearing system. Sec. 4 describes the case study, experimental setup, and the baseline method Sec. 5 presents the results of and offers a thorough discussion. Finally, Sec. 6 summarizes key findings and proposes directions for further research.

2. Virtual Sensor for Load Prediction

In this paper, we establish the notation where bold uppercase letters (e.g., $\mathbf{X}$), bold lowercase letters (e.g., $\mathbf{x}$), and calligraphic letters (e.g., $\mathcal{V}$) to denote matrices, vectors, and sets, respectively. Time steps are indicated by Superscripts (e.g., $\mathbf{X}^t$ is the matrix $\mathbf{X}$ at time t), while subscripts identify specific nodes (e.g., $\mathbf{x}_i$ is the vector for node i).

2.1. Problem Statement

In our case study, we focus on monitoring a bearing with a heterogeneous network of sensors. The data are collected from a test rig and comprise N sensor signals captured at discrete time instances. We particularly examine temperature and vibration data, which are represented as vectors:

$$\mathbf{x_T}^t = [x_{T_1}^t, x_{T_2}^t, ..., x_{T_{N_T}}^t]^T \in \mathbb{R}^{N_T}, \quad (1)$$

$$\mathbf{x_V}^t = [x_{V_1}^t, x_{V_2}^t, ..., x_{V_{N_V}}^t]^T \in \mathbb{R}^{N_V}, \quad (2)$$

where N_T and N_V are the number of each sensor type, while $x_{T_i}^t$ and $x_{V_i}^t$ denote the measurements at time t from the i^{th} sensor for temperature and vibration. Additionally, the rotational speed is recorded as $w^t \in \mathbb{R}$ at time t. Importantly, this characterizes the system's operational state and acts as a control parameter, rather than being a direct sensor measurement. To construct time-series samples for each sensor type, we employ a sliding window of length L, resulting in the following representations:

$$\mathbf{X_T}^{t_l:t} = [\mathbf{x_T}^{t_l}, \cdots, \mathbf{x_T}^{t-1}, \mathbf{x_T}^t] \in \mathbb{R}^{N_T \times L}, \quad (3)$$

$$\mathbf{X_V}^{t_l:t} = [\mathbf{x_V}^{t_l}, \cdots, \mathbf{x_V}^{t-1}, \mathbf{x_V}^t] \in \mathbb{R}^{N_V \times L}, \quad (4)$$

$$\mathbf{w}^{t_l:t} = [w^{t_l}, \cdots, w^{t-1}, w^t] \in \mathbb{R}^L, \quad (5)$$

where $t_l = t - L + 1 > 0$ marks the beginning of the observation window.

Our goal is to develop a function f, referred to as a virtual sensor, to accurately estimate the bearing load $\mathbf{y}^t \in \mathbb{R}^d$ at time t, targeting both axial and radial loads ($d = 2$). This function learns from heterogeneous sensor data $\mathbf{X_T}^{t_l:t}$, $\mathbf{X_V}^{t_l:t}$, and $\mathbf{W}^{t_l:t}$. Several challenges arise in developing such a function. Firstly, temperature and vibration signals exhibit inherently distinct characteristics. Temperature signals, typically monitored at lower frequencies, reflect gradual changes in the system's thermal state. In contrast, vibration signals are captured at high frequencies, offering insights into the immediate mechanical interactions and anomalies within the system. These differences in frequency not only affect the data processing strategy but also the interpretation of these signals in real-time monitoring. Additionally, the dynamic operating conditions introduce further complexity. Variations in load, speed, and environmental factors can significantly alter the base characteristics of both temperature and vibration data.

3. Graph-based Load Prediction Model

3.1. Framework Overview

We propose a novel Heterogeneous Temporal Graph Neural Network (HTGNN) for real-time bearing load prediction. Our framework learns a virtual sensor function, $f(\mathbf{X_T}^{t_l:t}, \mathbf{X_V}^{t_l:t}, \mathbf{W_T}^{t_l:t}) = \mathbf{Y}^t$, to accurately estimate the bearing load $\mathbf{Y}_t$ at a given time t. The HTGNN's main novelty lies in its ability to effectively capture the heterogeneity of sensor data and model the interactions between different sensor types. We achieve this by representing different sensor types as distinct node types in an aggregated temporal graph. This allows us to extract unique dynamics of each sensor type using tailored models and then model their interactions with specialized GNNs, This offers a significant advantage over traditional homogeneous temporal GNN methods that consider only a single type of relation. Fig. 1 illustrates the HTGNN architecture. The model's key components are:

1. **Heterogeneous temporal graph construction**, which constructs the bearing graph. (Sec. 3.2).
2. **Context-aware heterogeneous dynamics extraction**, which captures dynamics of different sensor types (Sec. 3.3).
3. **Heterogeneous interaction modelling**, which models complex interactions between diverse sensors (Sec. 3.4).
4. **Load prediction**, which predicts the bearing loads using the learned node representations (Sec. 3.5).

In the following, we detail each component of HTGNN.

3.2. Heterogeneous Temporal Graph Construction

Heterogeneous Static Graph. Following (Shi, 2022), a Heterogeneous Static Graph (HSG), denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consists of a node set $\mathcal{V}$ and an edge set $\mathcal{E}$, where nodes and edges can be of different types. The graph is associated with a node-type mapping function $\phi : \mathcal{V} \to \mathcal{A}$ and an edge-type mapping function $\psi : \mathcal{E} \to \mathcal{R}$, with $\mathcal{A}$ and $\mathcal{R}$ representing the sets of node and edge types, respectively, satisfying $|\mathcal{A}| + |\mathcal{R}| > 2$.

Heterogeneous Temporal Graph. Extending the concept of a Heterogeneous Static Graph (HSG), a Heterogeneous Temporal Graph (HTG) is defined as a sequence of HSGs over T time steps, $\mathcal{G}^T = \{\mathcal{G}^{t_1}, \ldots, \mathcal{G}^{t_T}\}$. Each graph $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$

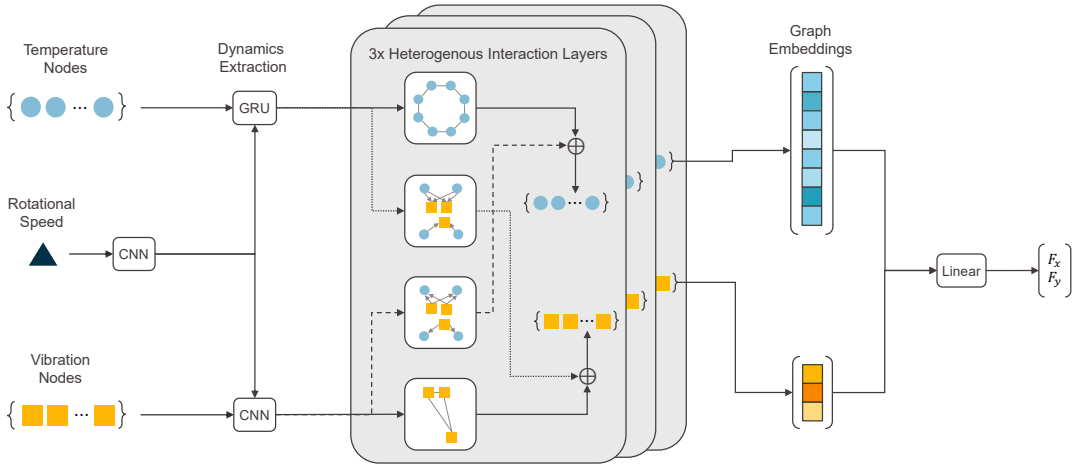


Figure 1. Architecture of the proposed Heterogeneous Temporal Graph Neural Network (HTGNN) for Load Prediction.

within this sequence represents the state of the graph at time t. The node and edge type mapping functions, ϕ and ψ remain consistent across time steps. The HTG can then be presented in an aggregated form as:

$$\mathcal{G}^T = \left(\bigcup_{t=t_1}^{t_T} \mathcal{V}^t, \bigcup_{t=t_1}^{t_T} \mathcal{E}^t \right), \tag{6}$$

combining nodes $\mathcal{V}^t$ and edges $\mathcal{E}^t$ across all time steps while preserving heterogeneity defined by ϕ and ψ.

Bearing graph construction. To model the heterogeneous sensor signals from a sensor network of the bearing system, we construct an HTG. This graph consists of two types of nodes: temperature (T) with attributes $\mathbf{X_T}^{t_l:t}$ and vibration (V) with attributes $\mathbf{X_V}^{t_l:t}$. Edge types represent relationships between node types: T-T, V-V, T-V, and V-T. We assume that these relationships are invariant over time. The HTG allows capturing the interactions and evolution of temperature and vibration signals within the bearing system. A visualization of the HTG is provided in Fig. 1.

3.3. Context-aware Node Dynamics Extraction

In complex systems, the behavior of individual nodes (sensors) is often influenced by the global operating context. In our bearing system, rotational speed can be considered a control variable, where increases in rotational speed lead to higher vibration intensity and faster temperature rises. To capture these important influences, our HTGNN model leverages context-aware dynamics extraction for node, following the strategy proposed in (Zhao & Fink, 2023). We extract contextual information from rotational speed and integrate it into the dynamics modeling of other sensor types using tailored techniques.

Rotational speed. To extract meaningful representations of operational state context from the rotational speed signal, which contains noise, we employ a 1D Convolutional Neural Network (1DCNN). We choose a 1DCNN due to its effectiveness in capturing patterns within time-series data. This process generates a hidden representation of dimensionality $\mathbf{h}_w \in \mathbb{R}^{d_w}$, which is used to augment the dynamics extraction from other sensor types. Our 1DCNN configuration adopts channel sizes [2, 2, 1], kernel sizes [3, 5, 5], and employs the SiLU activation function:

$$\mathbf{h}_w = \text{SiLU}\left(\text{1DCNN}(\mathbf{w}^{t_l:t})\right), \tag{7}$$

Temperature. We model the temperature dynamics using a Gated Recurrent Unit (GRU) network. For each temperature node j, the GRU updates its cell state at each time step τ to capture the temporal dynamics within the sequence $\mathbf{x_T}_j^{t_l:t}$. Importantly, we initialize the GRU's hidden state with $\mathbf{h}_w$ (rotational speed encoding from Eq. 7), allowing the operational state context to influence temperature dynamics:

$$\mathbf{h_T}_i^{\tau} = \text{SiLU}\left(\text{GRU-Cell}(\mathbf{x_T}_i^{\tau}, \mathbf{h_T}_i^{\tau-1})\right), \forall \tau \in [t_l, t]. \tag{8}$$

We use the final state $\mathbf{h_T}_i^{t} \in \mathbb{R}^{d_T}$, representing the encoded dynamics of node i up to time t and incorporating the operational state context, as the temperature node representation $\mathbf{h_T}_i \in \mathbb{R}^{d_T}$.

Vibration. Similar to the rotational speed encoding, we use a 1DCNN to model the dynamics of vibration signals. This process learns the hidden representation $\mathbf{h_V}_i^{t}$ from the vibration sequence $\mathbf{x_V}_i^{t_w:t}$ of a vibration signal i:

$$\mathbf{h_V}_i^{t} = \text{SiLU}\left(\text{1DCNN}(\mathbf{x_V}_i^{t_l:t})\right). \tag{9}$$

Finally, we concatenate $\mathbf{h_V}_i^{t} \in \mathbb{R}^{d_V}$ with $\mathbf{h_w} \in \mathbb{R}^{d_w}$ to form the complete node representation $\mathbf{h_V}_i = \left[\mathbf{h_V}_i^{t} \parallel \mathbf{h_w}\right] \in \mathbb{R}^{d_V+d_w}$. This incorporates both vibration dynamics and operational state.

3.4. Heterogeneous Interaction Modelling

We model heterogeneous interactions between different sensor types to capture the influence of operating context-aware dynamics. The proposed HTGNN model addresses two types of interactions within the graph: interactions among the same type of nodes and interactions across different types. This interaction modeling applies to node dynamics previously extracted in the node dynamics extraction section (temperature node from Eq. 8, vibration node from Eq. 9).

Same-type interactions. For interactions among nodes of the same sensor type, we employ Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017). This allows us to refine node representations by aggregating information from neighboring nodes that share similar characteristics. Messages passed from node j to node i of the same type with relation $r_s \in \mathcal{R}_\text{s}$ are computed as follows:

$$m_{j\to i}^{(l,r_s)} = \frac{1}{\sqrt{\hat{d}_i}\sqrt{\hat{d}_j}} \mathbf{W}_{\phi(j),r_s}^{(l)} \mathbf{h}_j^{(l)}, \forall r_s \in \mathcal{R}_\text{s}, \phi(j) = \phi(k), \tag{10}$$

where $\hat{d}_i$ and $\hat{d}_j$ denote normalized node degrees, and $\mathcal{R}_\text{s}$ is the set of edge types connecting nodes of the same type.

Different-type interactions. To model the influence of one sensor type on another (e.g., the impact of temperature on vibration), we utilize Graph Attention Networks v2 (GATv2) (Brody et al., 2022). This mechanism dynamically computes attention-weighted messages, allowing the model to discern the varying importance of different neighbors. The attention coefficients $\alpha_{jk}^{(l,r_d)}$ for a target node i receiving a message from node j with relation $r_d \in \mathcal{R}_d$ are defined as:

$$\alpha_{jk}^{(l,r_d)} = \text{softmax}_j \left(\mathbf{a}_{r_d}^{(l)T} \text{LeakyReLU}(\mathbf{W}_{r_d}^{(l)} \cdot [\mathbf{h}_i^{(l)} \parallel \mathbf{h}_j^{(l)}]) \right), \tag{11}$$

where $r_d \in \mathcal{R}_d$ represents the set of edge types connecting nodes of different types. Messages are then computed as:

$$m_{j\to i}^{(l,r_d)} = \alpha_{jk}^{(l,r_d)} \mathbf{W}_{\phi(j),r_d}^{(l)} \mathbf{h}_j^{(l)}, \forall r_d \in \mathcal{R}_\text{d}, \phi(j) \neq \phi(k), \tag{12}$$

Aggregation and update: After aggregating messages of both same-type and different-types, the node representations are updated as follows:

$$\mathbf{h}_{\phi(i)}^{(l+1)} = \text{SiLU} \left(\sum_{r \in \mathcal{R}_s \cup \mathcal{R}_d} \sum_{j \in \mathcal{N}_r(i)} m_{j\to i}^{(l,r)} \right). \tag{13}$$

3.5. Load Prediction

Having extracted the context-aware dynamics of each node, we now combine the heterogeneous node representations to learn the virtual sensor function $f(\mathbf{X_T}^{t_l:t}, \mathbf{X_V}^{t_l:t}, \mathbf{W_T}^{t_l:t}) = \mathbf{Y}^t$. We achieve this by flattening the final node representations into a unified input vector for a Multilayer Perceptron (MLP). The MLP processes this aggregated information and outputs two values: the predicted axial and radial loads.

To ensure the model's accuracy under real-world conditions, the training objective is to minimize the L1 loss between the predicted bearing load $\hat{\mathbf{y}}^i$ and the actual load $\mathbf{y}^i$. We choose L1 loss for its robustness to outliers. This is particularly important in bearing systems, occasional measurement noise or transitional operating conditions might generate extreme data points. The loss is defined as $\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} \left| \hat{\mathbf{y}}^i - \mathbf{y}^i \right|$, where M is the number of training samples.

4. Case Study

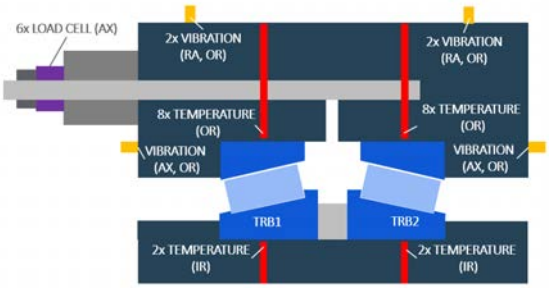


Figure 2. Cross-sectional view of the bearing test rig indicating sensor types and installation locations.

The data used in this study was collected at the SKF Sven Wingquist Test Centre (SWTC) using a face-to-face test rig with two identical single-row tapered roller bearings (TRBs). The TRBs feature a rotating inner ring, an outer diameter of 2,000 mm, an inner diameter of 1,500 mm, and a width of 220 mm, each incorporating 50 rollers. This setup aims to assess load conditions under various operational scenarios. Fig. 2 illustrates the sensor positioning on both identical TRBs. Ten temperature sensors are positioned on each bearing (eight uniformly distributed on the outer ring (OR), two on the inner ring (IR)). Additionally, six vibration sensors on the outer ring measure both axial (AX) and radial (RA) vibrations, with sensors placed at the top and bottom of the bearing housing for the radial direction.

Temperature is recorded at a 1 Hz sampling rate with a precision of 0.05°C. Vibration data is resampled to 1 Hz through RMS aggregations. Axial and radial forces are measured and controlled by several load cells, with an aggregated load value in both directions used as a ground truth for this study (note that the radial load cells are not shown in the figure).

4.1. Data Preprocessing

To reduce noise and transient fluctuations in the temperature data, we apply a moving average filter with a 1-minute window. We focus on the rate of temperature change because the bearing temperature responds gradually to changes in load and speed. We calculate this rate over 5-minute periods to

align with typical operational changes. This approach allows our model to identify the immediate impact of load changes on temperature, rather than the cumulative effects of historical variations. After preprocessing, we split both temperature and vibration signals using a sliding window, with a length of 30 seconds and a stride of 1 second.

4.2. Train-Test Split

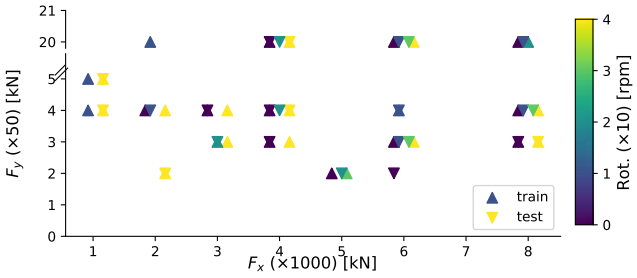


Figure 3. Train-test split of bearing load conditions for vibration data analysis (55% training, 45% testing)

We divided the dataset into training, validation, and testing sets. Approximately 55% of the data (924,230 samples across 31 unique operating conditions) was used for training and validation, with a random 80/20 split. The remaining 45% (699,340 samples across 25 unique operating conditions) was reserved for testing. We included only cases that maintained stationary operation for at least 10 minutes and up to 2 hours. We ensure that each case (a unique combination of axial load F_x, radial load F_y, and rotational speed) maintained stationary operation for at least 10 minutes and up to 2 hours. In total, the dataset comprised 56 unique operating conditions. To assess generalization, 12 conditions in the test set were unseen from the training and validation data. Fig. 3 provides a detailed breakdown of the specific conditions included in the training and test sets.

4.3. Heterogeneous Bearing Graph Construction

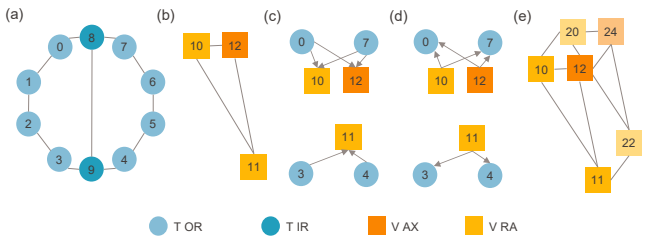


Figure 4. Heterogeneous graphs for bearing sensor network relationship modeling. (a) T-T (b) V-V (c) T-V (d) V-T (e) connectivity across two test rig bearings.

We construct a heterogeneous graph with nodes representing sensors (temperature (T) and vibration (V)). Temperature nodes are further classified into inner ring (T IR) or outer ring (T OR) nodes. V nodes, which are installed on the outer ring, are distinguished by their load direction: radial (V RA) or axial (V AX). We model four types of relationships: T-T, V-V, T-V, and V-T. Here, T-T and V-V represent homogeneous relationships, while V-T and T-V represent heterogeneous relationships. Node positions reflect physical sensor placement. Fig. 4(a) and (b) illustrate the connectivity within a single bearing based on physical proximity. Additionally, IR nodes are connected due to relatively uniform temperatures across the inner ring. Given that the test rig consists of two bearings, we connect them based on proximity, as illustrated in Fig. 4(e) for V nodes. We assume symmetrical (undirected) relationships within the same sensor type and model heterogeneous T-V and V-T relationships with directed edges, as demonstrated in Fig. 4(c) and (d).

4.4. Experimental Setup

Baseline. We employ a 1DCNN model as our baseline due to its established success in handling multivariate time series data. 1DCNNs are particularly well-suited for signal prediction tasks, making them a strong baseline. We adapt the design from (Chao et al., 2022), tailoring the architecture to our specific dataset through a grid search to minimize the mean absolute error (MAE) on the validation set. The explored parameter spaces included hidden channel dimension (20, 50, or 100), kernel size (3, 5, or 9), number of channels (20, 50, or 100), and number of layers (3, 4, or 5). The optimized model consists of four layers, each with 100 channels with 100 hidden dimensions, a kernel size of 9, batch normalization, a dropout rate of 0.5 for regularization, and a SiLU activation function (consistent with our proposed method). This configuration has a total of 209,403 parameters.

HTGNN hyperparameter tuning. We similarly used grid search for HTGNN hyperparameter tuning. To reduce the search space, we maintained a consistent hidden size across all layers and the same graph embedding dimension for all GNN modules. The search space comprised: node embedding dimension (values of 10, 15, 20), number of GNN layers (2 or 3), GNN hidden dimension (40 or 80), graph head hidden dimension (40 or 80), and number of graph head layers (2 or 3). The optimal HTGNN configuration consists of a node embedding dimension of 10, 3 GNN layers with a hidden dimension of 80, and a graph head dimension of 40. The configuration has a total of 142,394 parameters.

Training. We optimized the HTGNN and 1DCNN models using the AdamW optimizer with a learning rate of 1e-3. Training was continued for up to 50 epochs with early stopping at 30 epochs with patience of 10 steps. We used a batch size of 512 and minimized L1 loss (defined in Sec. 3.5). To ensure the robustness of our results, experiments were repeated five times with different initializations, and the mean and standard deviation of the results were reported.

Table 1. Averaged model performance over cases and runs

		1DCNN	HTGNN
Seen	MAE_{F_x} (kN)	531.3	**203.1**
	MAE_{F_y} (kN)	33.2	**12.4**
	$MAPE_{F_x}$ (%)	12.8	**4.5**
	$MAPE_{F_y}$ (%)	12.0	**5.7**
Unseen	MAE_{F_x} (kN)	1765.5	**1649.7**
	MAE_{F_y} (kN)	58.7	**57.4**
	$MAPE_{F_x}$ (%)	33.2	**29.2**
	$MAPE_{F_y}$ (%)	17.8	**15.8**

5. Results

We evaluate the model performance on Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

Superior performance on seen conditions. Tab. 1 highlights the HTGNN model's superior performance advantage compared to traditional 1DCNN models in predicting seen conditions. Notably, this improvement is evident in both axial (Fx) and radial (Fy) load predictions, with the HTGNN achieving approximately one-third the MAPE for Fx and half the MAPE for Fy compared to the 1DCNN. Importantly, the scenario considered here reflects real-world conditions. It is feasible for the sensor roller to collect load data across all typical operating conditions before its battery depletes, allowing the HTGNN to serve as a reliable virtual sensor.

HTGNN's physical prior. The superiority of the HTGNN in unseen conditions highlights the advantages of explicitly modeling heterogeneous sensor relationships. The physical connectivity in the bearing system acts as an effective inductive bias for the model. We attribute the improved performance of the HTGNN to its ability to capture complex interactions between temperature and vibration measurements, which often exhibit interdependent behaviors in bearing systems. The proposed architecture of the HTGNN is ideally suited to represent these heterogeneous relationships. In contrast, 1DCNN's homogeneous approach to processing variables limits its ability to model such complex interdependencies, leading to higher prediction errors.

Better generalizability. Fig. 6 presents the mean MAPE in F_x and F_y for various bearing load conditions, with unseen conditions highlighted in gray. Although the HTGNN generally outperforms the CNN in handling unseen conditions, as detailed in Tab. 1, there are instances depicted in Fig. 6 where CNN shows competitive performance. This challenge in generalization can be partially attributed to the dynamics shown in Fig. 5, which illustrates the significant effects of rotational speed changes on both vibration intensity and the rate of temperature change. Additionally, the underrepresentation of certain rotational speeds in the training data may impede interpolation, impacting the generalization capabilities of both models. Interestingly, as depicted in Fig. 5, the

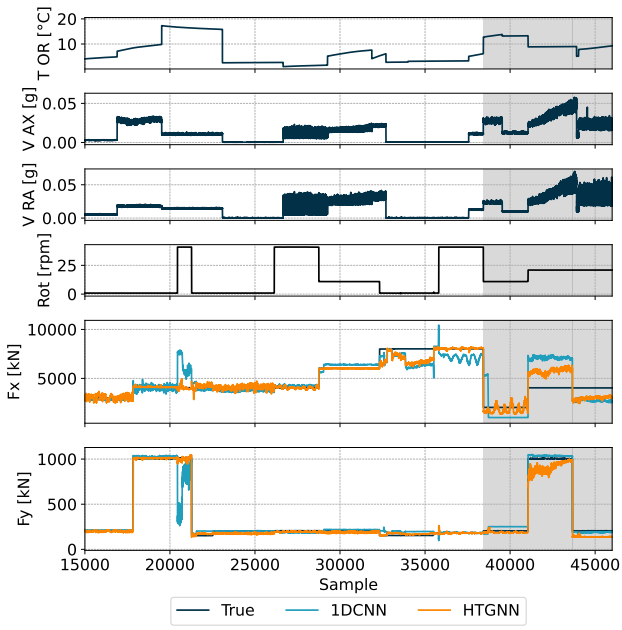


Figure 5. Examples of input signals and load prediction performance. Shaded areas indicate unseen conditions.

HTGNN tends to generalize better for the commonly encountered unseen rotational speed of 10 rpm than for the less frequently occurring speed of 20 rpm, across both axial and radial loads. For details on the distribution of conditions in the training and testing sets, see Fig. 3.

6. Conclusion

In this research, we propose HTGNN, a novel virtual sensor that accurately maps vibration and temperature signals under varying rotational speeds to axial and radial bearing load predictions. Our findings demonstrate that HTGNN outperforms 1DCNN models, particularly when trained on representative conditions. The success of HTGNN highlights the importance of incorporating physical priors and inductive biases: by modeling the connectivity of the bearing sensor network, HTGNN effectively captures the complex interactions between temperature and vibration. This superior performance suggests HTGNN's potential as a reliable virtual sensor in real-world applications, replacing battery-powered load sensors after their lifespan. This could facilitate proactive maintenance, reducing unexpected breakdowns and optimizing the lifespan of bearings. However, the models cannot generalize as effectively to unseen speed conditions. Future work should focus on investigating datasets that include a broader range of speed conditions in the training to determine if the model can improve its generalization capabilities. Additionally, measuring the model's performance using real load data measured from sensor rollers in real operations and not just from the test rig would be valuable.

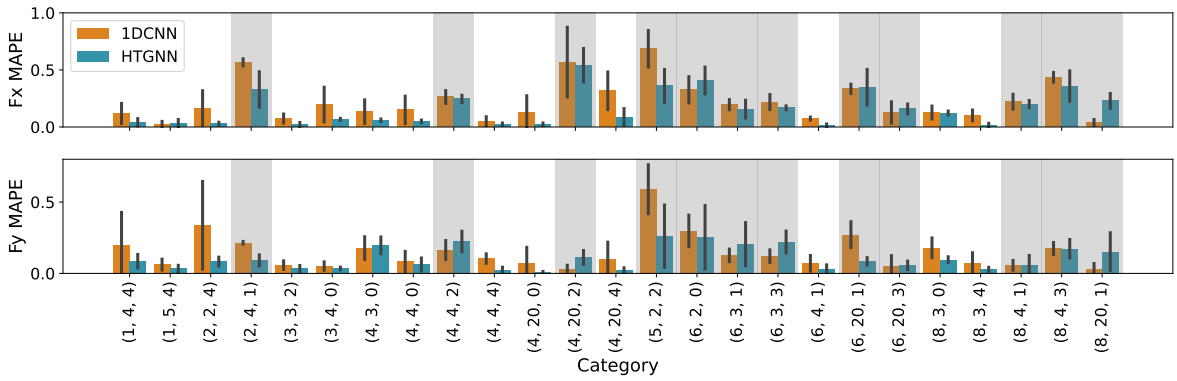


Figure 6. Mean test set performance averaged over 5 runs of CNN and HTGNN on bearing load conditions $F_x(\times 1000)$ [kN], $F_y(\times 50)$ [kN], and rotational speed $(\times 10)$ [r/min]. Shaded areas indicate unseen conditions.

REFERENCES

Baggerohr, S. (2023). On the detection of rolling contact fatigue in large bearings using roller embedded sensors. In *Surveillance, vibrations, shock and noise.*

Bilbao, J., Lourens, E.-M., Schulze, A., & Ziegler, L. (2022). Virtual sensing in an onshore wind turbine tower using a gaussian process latent force model. *Data-Centric Engineering*, *3*, e35.

Brody, S., Alon, U., & Yahav, E. (2022). How attentive are graph attention networks? In *International conference on learning representations (iclr).*

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 4027–4035).

Dimitrov, N., & Göçmen, T. (2022). Virtual sensors for wind turbines with machine learning-based time series models. *Wind Energy*, *25*(9), 1626–1645.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International conference on machine learning* (pp. 1263–1272).

Harris, T. A., & Kotzalas, M. N. (2006). *Essential concepts of bearing technology*. CRC press.

Hou, Y., & Wang, X. (2021). Measurement of load distribution in a cylindrical roller bearing with an instrumented housing: Finite element validation and experimental study. *Tribology International*, *155*, 106785.

Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., … Pan, S. (2023). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *arXiv preprint arXiv:2307.03759*.

Kerst, S., Shyrokau, B., & Holweg, E. (2019). A model-based approach for the estimation of bearing forces and moments using outer ring deformation. *IEEE Transactions on Industrial Electronics*, *67*(1), 461–470.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th international conference on learning representations.*

Konopka, D., Steppeler, T., Ottermann, R., Pape, F., Dencker, F., Poll, G., & Wurz, M. (2023). *Advancements in monitoring of tribological stress in bearings using thin-film strain gauges.*

Martin, D., Kühl, N., & Satzger, G. (2021). Virtual sensors. *Business & Information Systems Engineering*, *63*, 315–323.

Morales, G. E., Engelen, P., & Van Nijen, G. (2019). Propagation of large spalls in rolling bearings. *Tribology Online*, *14*(5), 254-266. doi: 10.2474/trol.14.254

Niresi, K. F., Zhao, M., Bissig, H., Baumann, H., & Fink, O. (2023). Spatial-temporal graph attention fuser for calibration in iot air pollution monitoring systems. In *2023 ieee sensors* (pp. 01–04).

Peng, D., Wang, H., Liu, Z., Zhang, W., Zuo, M. J., & Chen, J. (2020). Multibranch and multiscale cnn for fault diagnosis of wheelset bearings under strong noise and variable load condition. *IEEE Transactions on Industrial Informatics*, *16*(7), 4949–4960.

Shi, C. (2022). Heterogeneous graph neural networks. *Graph Neural Networks: Foundations, Frontiers, and Applications*, 351–369.

Song, Z., Hackl, C. M., Anand, A., Thommessen, A., Petzschmann, J., Kamel, O., … Hauptmann, S. (2023). Digital twins for the future power system: An overview and a future perspective. *Sustainability*, *15*(6), 5259.

Wang, G., Jia, Q.-S., Zhou, M., Bi, J., & Qiao, J. (2021). Soft-sensing of wastewater treatment process via deep belief network with event-triggered learning. *Neurocomputing*, *436*, 103–113.

Widner, R., & Littmann, W. (1976). Bearing damage analysis. *National Bureau of Standard special publication*(423), 1.

Yang, X., Zheng, Y., Zhang, Y., Wong, D. S.-H., & Yang, W. (2022). Bearing remaining useful life prediction based on regression shapalet and graph neural network. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–12.

Zhao, M., & Fink, O. (2023). Dyedgegat: Dynamic edge via graph attention for early fault detection in iiot systems. *arXiv preprint arXiv:2307.03761*.

A novel prognostics solution for accurate identification of degradation patterns in turbo machines with variable observation window

Unnat Mankad[1], Gabriele Mordacci[2], Aidil Fazlina Hasbullah[3], Fahzziramika Nadia Jaafar[4], Carmine Allegorico[5], and Gionata Ruggiero[6]

[1]*Baker Hughes, Bengaluru, 560037, India*
unnat.mankad@bakerhughes.com

[2,3,4,6] *Baker Hughes, Kuala Lumpur, 50400, Malaysia*
gabriele.mordacci@bakerhughes.com
aidilfazlina.hasbullah@bakerhughes.com
fahzziramikanadia.jaafar@bakerhughes.com
gionata.ruggiero@bakerhughes.com

[5]*Baker Hughes, Florence, 50127, Italy*
carmine.allegorico@bakerhughes.com

ABSTRACT

The degradation of a system is a time bound phenomenon, which leads to the deterioration of turbomachinery, in terms of performance and reliability. If undetected and not acted upon in time, this could also lead to sudden system failure, resulting in unplanned unit downtime and maintenance. Unplanned downtime of a turbomachine leads to severe production loss for the end customer and consequent economic damages. Early detection of a degradation pattern would provide the customer with the opportunity to timely carry out corrective actions, preventing an unscheduled down time. The paper evaluates degradation identification methodology currently known from literature and finds them not accurate enough for general purpose application required by the solution. The paper discusses a novel methodology which can accurately detect degradation patterns of timeseries data. Critical features of this methodology are novel time-based correlation enabled regression model with variable observation window, autonomous training, and automatic adjusting capability to incorporate operating behavior change or physical system replacement. This leads to high accuracy, high generalization, and domain agnostic application capability. Moreover, particular focus is given to achieving high probability of detection and a low probability of false alarm. The paper demonstrates the performance achieved by the methodology when applied to the field of prognostics and diagnostics of IoT connected turbomachines through 50+ real application cases.

1. INTRODUCTION

Rotating Turbomachines play a critical role in Industrial domain in Oil & Gas / Energy Plants serving various applications, such as Liquified Natural Gas (LNG), pipeline, fertilizers, refineries and power generation units. One of the most important aspects for the operators of these turbomachines are continuous availability and reduced downtime covering the entire life cycle. Iannitelli et al. (2018) highlighted that unscheduled shutdown of the turbomachines can have impact on the whole plant downtime with associated significant loss of production.

Baker Hughes is a leading Original Equipment Manufacturer of Rotating Turbomachines with a wide Product Range of Gas Turbines, Centrifugal Compressors, Pumps, Steam Turbines, Electric Motors, Axial Compressors, etc. These products have been operating in various Oil & Gas and Power Generation facilities around the globe covering all the segments of the entire value chain of Oil & Gas industry and have an unparallel operating history.

Baker Hughes has developed monitoring capabilities which are offered as a service, applied to a broad installed fleet of rotating equipment including gas turbines. Baker Hughes' iCenter ecosystem continuously acquires different sensor parameters of its deployed assets at customer premises. These large number of operational data from the everyday operation of turbomachines is usually collected and analyzed by means

of analytics, component models and rules implemented by subject matter experts, as soon as new data is transferred to the monitoring center. Allegorico & Mantini (2014) indicated that anomaly detection rules and models are designed to scan through the data and notify the monitoring and diagnostic engineers, if any anomalies or emerging problems are detected. All alerts are analyzed by diagnostic engineers along with trouble shooting analysis and useful insights are sent to customers comprising set of recommendations.

With turbomachinery covering various applications and operating in different operational scenarios, Baker Hughes follows a hybrid approach consisting of physics and data driven methods, where strong OEM knowledge is further enhanced by state-of-the-art data science methodologies to create robust solutions. This approach can be applied to the entire fleet of operating machines, to bring economies of scale and help maximize the availability and uptime of the monitored units.

1.1. Degradation phenomena

In Turbomachines, degradation phenomena accumulate over a period of time. Zagorowska, et al. (2019) indicated that degradation in turbomachine is an unwarranted phenomenon which deviates from the expected behavior and that changes the behavior of the affected system. Few examples of degradation include clogging of filters, performance degradation of compressors, increase spread of exhaust temperature measurement of gas turbines, etc. If degradation is not detected early, this may lead to a gradual build up above the mechanical integrity of the system which can cause sudden failures, break down and consequent downtime of the turbomachine with production loss for the end customer. A typical example of degradation in turbomachinery systems concerns filters. A filter acts as a mechanical stop for contaminants, to make sure they do not pass through the downstream systems. Due to their nature, filters have a tendency to get clogged or choked after a period of operation with gradual buildup of contaminants, creating a higher resistance to the flow. To detect abnormal operating conditions, analytics could be built to observe the behavior of the component by monitoring physical quantities, such as the pressure drop on the filter. This can be analyzed to infer information on its actual defect state. The ability to promptly detect these deteriorating conditions could be useful for implementing corrective actions.

Generally, degradation phenomena cannot always be directly measured, however it is possible to make use of indirect information or calculated parameters to verify the level of degradation of a system (for example the level of fouling of an axial compressor can be determined indirectly through the analysis of its compression efficiency). In general, the presence of a degradation phenomenon is signaled by the fact that the timeseries of interest shows a drift over time. If the timeseries has an upward trend, it is considered a positive degradation, otherwise it is considered a negative degradation.

In the current study, authors have focused on univariate time series with a stationary behavior in the normal operating range of the system. In these cases, a monotonic signal trend is considered anomalous and possibly linked to an ongoing degradation phenomenon. In the event that this monotonic trend is accompanied by a similar behavior of other signals related to it, the event is considered non-independent and therefore not anomalous.

Figure 1 shows a typical behavior of a sensor going through a degradation trend. As the sensor value increases over a period, this is considered a positive degradation phenomena.

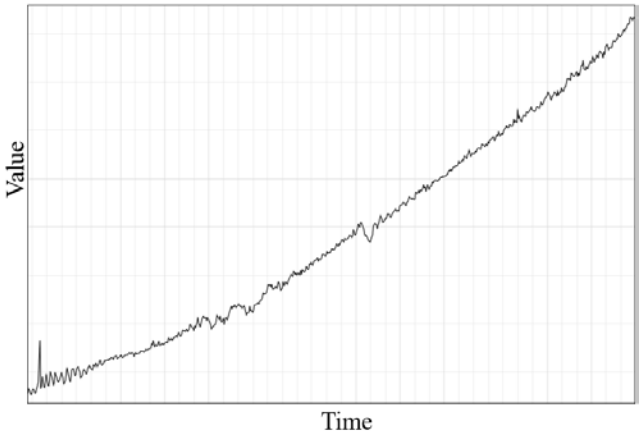


Figure 1. Example of a degradation pattern in a generic signal

In the analysis of degradation phenomena, another factor to consider is the observation time window. Figure 2 highlights the behavior of the same signal over a longer observation time.

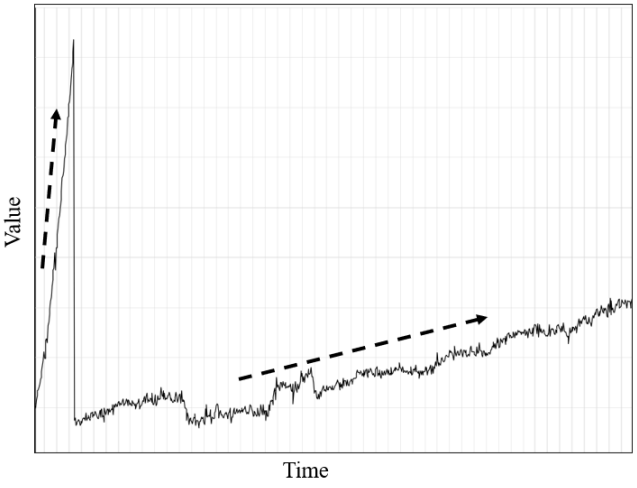


Figure 2. Example of multiple degradation patterns for the same signal

In this case, two different degradation profiles can be noted that evolve with two different time scales. The first degradation profile is quite sudden, while the subsequent one is rather slow and it build up over a longer period of time. The sudden drop down of the signal after the first degradation pattern is currently excluded from the current analysis.

2. EXISTING METHODS FOR TREND IDENTIFICATION

2.1. Monotonicity Trend – Mann-Kendall Test

The purpose of the Mann-Kendall (MK) test (Mann 1945, Kendall 1975, Gilbert 1987) is to statistically assess if there is a monotonic upward or downward trend of the variable of interest over time. A monotonic upward or downward trend means that the variable consistently increases or decreases through time.

The MK method calculates test statistics as the count of positive and negative deltas in the dataset.

$$\boldsymbol{S} = \sum_{\boldsymbol{j-1}}^{\boldsymbol{n-1}} \sum_{\boldsymbol{i-j+1}}^{\boldsymbol{n}} \boldsymbol{sgn}(\boldsymbol{x_i} - \boldsymbol{x_j}) \quad (1)$$

Where x is the observation value, i and j are time indices.

If number of observations, $n \geq 10$, Variance of S is calculated as follows.

$$VAR(S)_{MK} = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p-1}^{g} t_p\,(t_p-1)(2t_p+5)\right] \quad (2)$$

where g is the number of clusters of data points having the same data value and t_p is the number of observations in the p th group.

For example, in the sequence of observation in time $\{28, 32, 34, 2, 29, 32, 2, 34, 32\}$ there are $g = 3$ tied groups. Tied group $t_1 = 2$ for tied value of 2, tied group $t_2 = 3$ for tied value of 32 and tied group $t_3 = 2$ for tied value of 34

MK Test statistics is calculated as follows:

$$\begin{aligned} Z_{MK} &= \frac{S-1}{\sqrt{VAR(S)_{MK}}} \; if\; S > 0 \\ Z_{MK} &= 0 \; if\; S = 0 \\ Z_{MK} &= \frac{S+1}{\sqrt{VAR(S)_{MK}}} \; if\; S < 0 \end{aligned} \quad (3)$$

A positive value of Z_{MK} indicates an increasing trend, while a negative value of Z_{MK} indicates a decreasing trend.

The MK test was applied on the generic signal in Figure 3, which shows a clear degradation trend in different periods of time. The points where the increasing trend is detected are highlighted in orange.

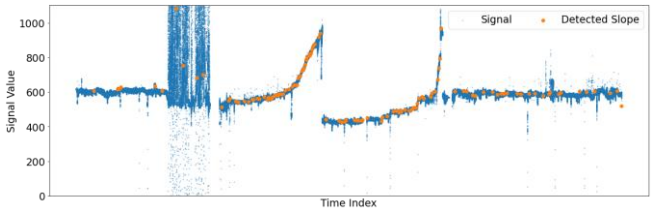


Figure 3. MK Test results

As observed from the Figure 3 for the given data set, the method was promising in terms of detecting the slope region, however it produced many false positives. To improve this manual threshold tuning is required, however this is not a practical and most effective solution for more general and scalable applicability of the methodology.

2.2. Theil Sen Slope Method

Theil (1950) proposed the median of pairwise slopes as an estimator of the slope parameters. Sen (1968) extended this estimator to handle ties. Sprent et al, (1993) indicated that Theil-Sen estimator is a regression method, robust to outliers.

Thiel-Sen estimator calculates the slope by taking the median of the slopes between each pair of points in the data. For a pair of points, $(x_i,\ y_i)$, *the slope is calculated as*

$$slope = \frac{(y_j - y_i)}{(x_j - x_i)} \quad (4)$$

An intercept between each pair of points, can be calculated as

$$b_i = y_i - m * x_i \quad (5)$$

where m is the Thiel-Sen slope. Following the similar methodology of finding the median of each slope between each pair of points, median of intercept is calculated.

Theil-Sen Slope method was applied on the same signal of Figure 3 after setting an appropriate threshold to detect the degradation pattern. Figure 4 shows the results of the Thiel Sen slope method.

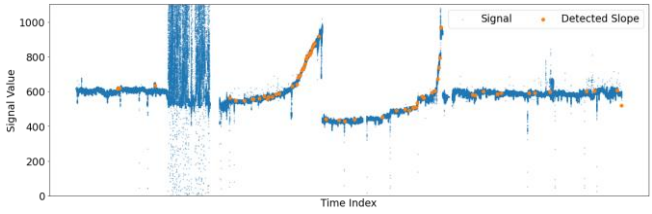


Figure 4. TS Test results

As observed from Figure 4 for given data set, Thiel-Sen method identifies trending patterns, however it still generates many false positives and is strongly dependent on the threshold value, which need to be manually adjusted based on the profile of the signal. This limits the general purpose and scalable applicability of the methodology.

3. NOVEL METHODOLOGY

Abernathy et al. (1973) indicated that sensor measurement are affected by noise and noise increase over a period of time as the sensor ages. Noise of the sensor measurement impacts the method development and it's ability to identify the degradation patterns. Furthermore, the degradation detection method must be easily scalable to other use cases and be able to work with different degradation patterns such as slow and fast degradations and presence of noise.

De Giorgi et al. (2023) have done an exhaustive literature review on detecting degradation phenomena as part of prognostic and diagnostics for jet engine health monitoring and have found that current literature degradation health monitoring techniques have certain gaps in terms of lack of standardization, lack of real world testing/comparative studies and limited consideration of multiple degradations.

Following the above analysis, it was concluded that current methods available in the literature may not effectively provide a generalized and robust solution. Furthermore, the existing methods are quite difficult to be fine-tuned in real application scenarios and are prone to generate a high rate of false positives.

As seen in Figure 3 & Figure 4, the degradation profile of a signal is a function of time. This could be caused by various factors, such as the intrinsic structure of the system, external interferences, natural aging and so on. In order to effectively capture degradation phenomena which evolves over a different time scale, authors had decided to distinguish 2 types of degradation profiles:

- Fast Degradation – These degradation profiles are quick with respect to typical behavior of the given signal/system.
- Slow Degradation – These degradation profiles slowly build over a period of time and may not show an obvious degradation behavior when the observation window is small.

Authors have then devised a novel methodology by filtering the signal into a High Frequency component and a Low Frequency component.

The High Frequency component of the signal is calculated as:

$$High\ Frequency_t = High\ Frequency_{t-1} * \alpha_t + Signal_t * (1 - \alpha_t) \quad (6)$$

Where α is the exponential smoothing average constant. As degradation phenomena are function of time and depends on past values, this constant has been selected to keep a balance between past observations and current values. After a careful analysis and various tests on real cases, this value was kept at 0.35.

The low frequency component of the signal is then calculated as

$$Low\ Frequency_t = Signal_t - High\ Frequency_{t-1} \quad (7)$$

Figure 5 shows the original signal and decomposition of the same into high & low frequency component of the given signal. Observing Low Filter, it is evident that, this features carries out the denoising of the signal.

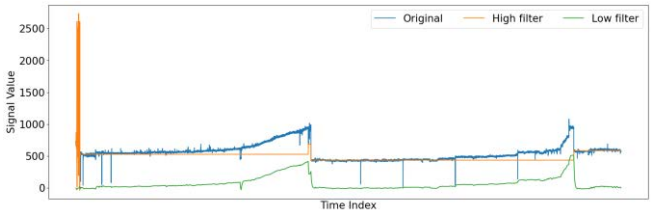


Figure 5. Signal decomposition in Low frequency

A novel time-based correlation approach was used to identify the degradation patterns of the low frequency component. The approach was based on the observation that if the signal is trending up or down over a period of time, it shall have a strong correlation with time, which will be positive or negative respectively.

The correlation coefficient was obtained by normalizing the covariance of the low frequency signal.

The covariance of the signal is calculated as

$$Covariance = E[XY] - (EX)(EY) \quad (8)$$

The variance is calculated as

$$Var_X = E[X^2] - E[X]^2$$
$$Var_Y = E[Y^2] - E[Y]^2 \quad (9)$$

Then the correlation coefficient is calculated as

$$Corelation\ Coefficient = \frac{Covariance}{\sqrt{Var_X * Var_Y}} \quad (10)$$

A threshold of 0.9 was then applied to this correlation coefficient to detect sections of the signal with high slope, as showed in Figure 6.

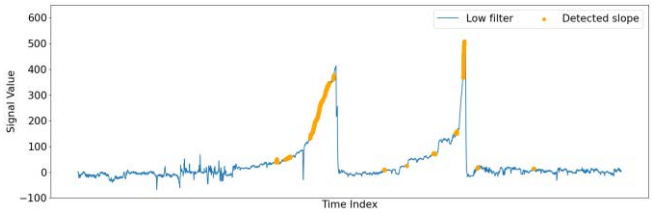


Figure 6. Time based correlation on Low frequency

As observed from Figure 6, the method is more robust and less sensitive if compared to previously discussed methods. It effectively captures the sections with high slopes; however, it is not capable of capturing the areas where degradation is slowest and it also generates some sporadic false alarms.

To overcome the limitation of the current method on the slow degradation patterns, the authors devised the dedicated approach described in the next paragraph.

3.1. Methodology for Slow Degradation

Verbai et al. (2024) applied linear regression method to identify and predict the degradation phenomena. Authors have further used the linear regression method to develop the methodology to capture slow degradation

The linear regression model is expressed as:

$$\hat{y}_i = b_0 + b_1 * x_i \quad (11)$$

Where $\hat{y}_i$ is the predicted value, b_0 is the intercept of the line, b_1 is the slope of the line, and x_i is the actual value.

The linear regression model is fit on 1 week of Low frequency data of the signal and further analysis is carried out on the line slope b_1, R^2 error and Root Mean Square Error.

R-squared (R^2) of the linear regression model is calculated as

$$R^2 = 1 - \left(\frac{SS_{residual}}{SS_{total}}\right) \quad (12)$$

Where $SS_{residual}$ is the sum of squares of the residual errors and SS_{total} is the total sum of the errors.

R^2 indicates the proportion of data points which lie within the line created by the regression model. A higher value of R^2 is desirable as it indicates a better fit.

To ensure a good regression model for subsequent analyses, a minimum value of R^2 score is required.

The Root Mean Square Error (RMSE) indicates the quality of predictions. It evaluates how far predictions are from the measured true values using Euclidean distance.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (13)$$

Where $\hat{y}_i$ is predicted value, y_i is actual true value and n is number of observations.

Above regression methodology was applied on low frequency of 7 days data. However in order to early detect the degradation phenomena, observation window considered was 1 day.

To further make sure that generated errors are within the typical operating range of the signal a threshold was applied on RMSE as a function of normal operating range of the signal.

To make sure, that only important degradation patterns are captured, a minimum threshold value was applied on the slope on top of already discussed threshold on R^2 and RMSE values. The proper value of the threshold was selected while doing an exhaustive testing to obtain a balance between False Positive and False Negative.

Results of Figure 7 shows the degradation pattern captured by the new methodology with high accuracy.

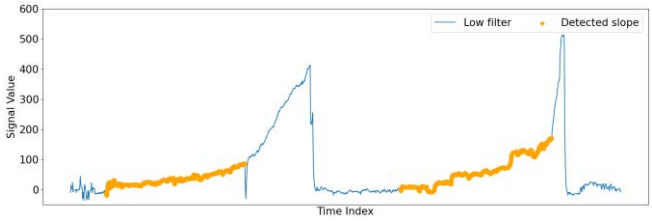


Figure 7. Detection by Slow Degradation Methodology

3.2. Time Window for Fast and Slow Degradation

Based on the extensive tests carried out and iteratively optimized, authors have specified 2 different observation periods, 24 hours and 7 days, which have proven to be effective on use cases that are common in our industry.

Table 1. Time Window duration

Degradation Type	Time Window
Fast Degradation	24 Hours
Slow Degradation	7 Days

3.3. Autonomous Training and Self Adjusting Capability

The developed methodology is intended to have a general-purpose application by covering various types of signals that are typically acquired on turbomachines. Moreover, it must

be able to function correctly for signals that span in a wide operating range.

To meet the above requirement, authors have developed a methodology to characterize the "Normal State" of operation of a signal, namely its typical operating range known from the past.

In order to define an operation range of the signal and detection of potential anomalous behavior of the signal, authors have then developed a typical operating range of signal called as confidence band.

Confidence Band is a function of the following signal statistical indicators and is calculated dynamically:

$$Confidence\ Band = f(Mean, Standard\ Deviation, Quantiles) \quad (13)$$

Any sustained operation outside of the normal state could be considered as a potential degradation pattern.

The method continuously updates the above statistics and redefines the system normal state when needed. Other factors that influence the signal behavior are the maintenance events such as major inspections, repairs, replacements, etc. and other external contributors like the process load and ambient conditions, which can lead to different operating behavior of a given signal. The algorithm is designed to self-adjust when this change in signal behavior occurs.

4. TECHNICAL CASES

Authors have extensively applied and tested this methodology on a variety of turbomachinery signals acquired by Baker Hughes' monitoring service. In the following section the authors reported some examples of real degradation events captured by applying this methodology on historical data. If not detected promptly, the progression of the degradation phenomenon could have caused the signal of interest to reach protection thresholds, causing alarms or even the trip of the unit. A trip leads to unavailability of the turbomachine and the loss of production for the end customer, with consequent economic damage.

The implemented methodology provides early detection of degradation of critical signals and provides the opportunity to perform corrective actions and increase the availability of turbomachinery.

This section captures few of the real technical cases captured from variety of signals acquired by Baker Hughes' monitoring service. Few of these signals are part of Centrifugal Compressors Auxiliary systems, Gas Turbines, etc. Some of the examples of these signals are Filter Differential Pressure, Vent Pressure, Compressor Efficiency etc,. As discussed before, these signals are expected to be stationary with in the normal operating range of the system. Any independent monotonic trend identification is considered to be anomalous behavior of the signal.

The grey are highlighted in the figures represents Confidence Band of the signal, which is the expected range of operation. As discussed before, methodology keeps on dynamically calculate this confidence band. Anomaly events are generated when the signal exceeds this confidence band.

4.1. Example of Fast Degradation

This section describes the example in which underlying degradation phenomena is Fast in nature and happens with in time window of 24 hours.

4.1.1. Fast Degradation Profile 1

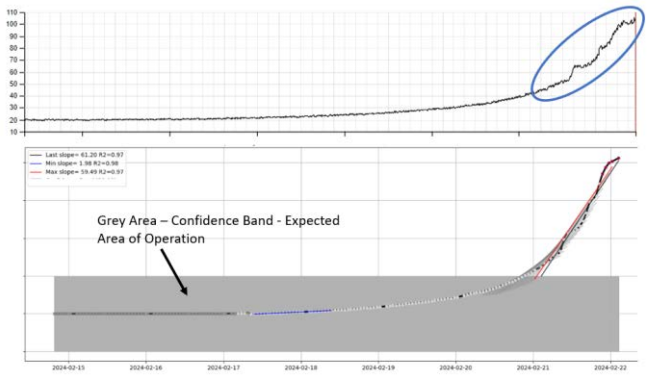


Figure 7. Detection of Fast Degradation event 1

4.1.2. Fast Degradation Profile 2

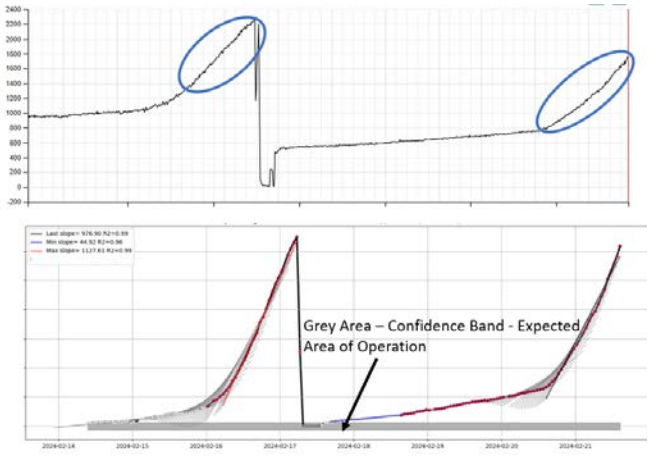


Figure 8. Detection of Fast Degradation event 2

As seen from Figure 7 & Figure 8, the methodology can effectively capture fast degradation of the signal, even when signal has high oscillation or a reset.

4.2. Example of Slow Degradation

This section describes the example in which underlying degradation phenomena is Slow in nature, accumulates over a longer period of time and happens with in time window of 7 days.

4.2.1. Slow Degradation Profile 1

Figure 9. Detection of Slow Degradation event 1

4.2.2. Slow Degradation Profile 2

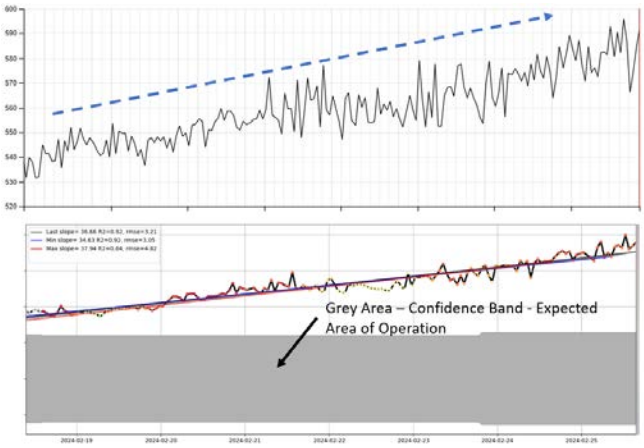


Figure 10. Detection of Slow Degradation event 2 – Noisy signal

In some cases, the monitored signal can be noisy and this may impact the detection capability of the algorithm. A such example is visible in Figure 10, where the raw signal is noisy, but at the same time shoes a slow degradation process. With the novel approach of the methodology, segregating low frequency of the signal, the methodology is effectively able to denoise the signal and accurately captures the degradation trend.

4.3. Timely Corrective Actions

The degradation patterns detected by this analytic could be associated to some specific failure modes of the system, thus mapping of this potential root cause with detected type of degradation phenomena is of high importance. Based on strong OEM knowledge, Baker Hughes has identified up to 8 root causes for degradation patterns. Some of these root causes are Instrument deviations, Clogging, Condensation, Process fluctuations, Fouling etc. With the given identified root cause, diagnostic engineers then propose a targeted corrective actions to the site service engineers. Implementation of this corrective actions eventually leads to improved uptime of the unit with no unscheduled shutdowns/repairs for the end customer.

5. Result Analysis

To summarize, the novel methodology proposed by the authors, separates High frequency and Low frequency component of the signal to effectively denoise the data and separate the rapid changes happening into the signal.

As the degradation profile is strongly dependent on the time interval, currently 2 observation windows have been considered. Results have shown that method is effectively able to capture the Fast and Slow degradation of the signal, whereas standard methods like Mann Kendall and Thiel Sen slope has not been very effective and accurate in either identifying the degradation trend or wrongly capturing the degradation. It is to be further noted that, analytic has quite good generalization capability as it is able to catch wide operating range of signal as observed from Figure 7, 8, 9 & 10.

In order to validate the methodology on a larger data set, the approach was applied on 600+ turbomachines being monitored by Baker Hughes's iCenter eco system. With extensive understanding of Turbomachines system, signals for validation were selected in such a way that signal show a degradation trend due to inherent malfunctioning of the system. Some of the examples of these signals are Filter Differential Pressure, Vent Pressure, etc. Methodology was tested in a Batch process where incoming data with a given sampling frequency of 1 minute was processed in a batch of 2 hours.

To calculate the key performance indicators of the methodology, a manual approach was used which required a great effort from the subject matter experts to analyze all the events generated by the algorithm. The methodology was also tested on a number of real cases of degradation that were already known to the monitoring service.

Table 2 shows the performance metrics of the method implemented.

Table 2. Performance of Method during Validation

Details	Value
Number of Assets on which methodology was applied	600+
Average processing time for 2 hours batch with 1 minute sampling	1.1 seconds / asset
Total Degradation events captured on multiple signals	50+
Probability of Detection	> 95%

False Positive Rate	< 5%
False Negative Rate	< 5%
Precision	> 95%
Recall	> 95%

6. CONCLUSION

In this paper, the authors discussed the problem of detecting degradation phenomena in the application field of turbomachinery and explained the importance of implementing early detection of such events in Baker Hughes continuous monitoring service.

Authors have also described degradation phenomena which accumulates with time scales of different duration, happening on different signals acquired on turbomachines. The existing methods for the identification of degradation patterns, already known in the literature, have not been deemed accurate enough for general purpose applicability required by the solution. A novel approach has been developed comprising strong features, like the extraction of low frequency component of the signal, the incorporation of time based correlation and linear regression model applied on multi time observation window. It was shown that these unique features empower the method with accurate detection rate, precision and recall. The proposed methodology also embeds autonomous learning and auto setting capability that enables generalized application covering multiple types of signals with wide operating ranges.

To validate the new methodology on a large data set, tests were performed on historical timeseries data from more than 600+ turbomachines being monitored by Baker Hughes's iCenter eco system. The signals were chosen on some families of mechanical systems which generally can present degradation phenomena during their life cycle. The paper then also discusses some real detection cases and explains the process through which the probable associated root causes are identified and the corrective actions are suggested to the final customers for field implementation. Finally, the performance matrix of the methodology is shown, which was found to comply with the stringent detection requirements followed by Baker Hughes.

NOMENCLATURE

LNG	Liquified Natural Gas
OEM	Original Equipment Manufacturer
MK	Mann Kendall
TS	Thiel Sen
RMSE	Root Mean Square Error

REFERENCES

Abernathy, R. B., Powell, B. D., Colbert, D. L., Sanders, D. G., & Thompson Jr., J. W. (1973). *Handbook Uncertainty in Gas Turbine Measurements*. Fort Belvoir, VA, USA : DTIC.

Allegorico, C., & Mantini, V. (2014). A Data-Driven Approach for on-line Gas Turbine Combustion Monitoring using Classification Models. *Proceedings of PHM Society European Conference*. July 8-10, Nantes, France. https://doi.org/10.36001/phme.2014.v2i1.1461

De Giorgi, M.G., Menga, N., & Ficarella, A. (2023). Exploring Prognostic and Diagnostic Techniques for Jet Engine Health Monitoring: A Review of Degradation Mechanisms and Advanced Prediction Strategies. *Energies*, vol. 16. https:// doi.org/10.3390/en16062711

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. New York: Wiley.

Iannitelli, M., Allegorico, C., Garau, F., & Capanni, M. (2018). A hybrid model for on-line detection of gas turbine lean blowout events. Proceedings of PHM Society European Conference. July 3-6, Utrecht, The Netherlands. https://doi.org/10.36001/phme.2018.v4i1.405

Kendall, M.G. (1975). *Rank Correlation Methods*. London: Charles Griffin.

Mann, H.B. (1945). Non-parametric tests against trend, *Econometrica*, vol. 13, pp. 163-171. doi:10.2307/1907187

Sprent, P. (1987). *Applied nonparametric statistical methods. tatistical Methods.* New York: CRC Press.

Sen, P. (1968). Estimated of the regression coefficient based on Kendall's Tau. *Journal of American Statistical Institute*, vol. 63, pp. 1379-1389

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, 1-2; confidence regions for the parameters of linear regression equations in two, three and more variables. *Indagationes Mathematicae ,* vol. 1, issue 2

Varbai, B., Wéber, R., Farkas, B., Danyi, P., Krójer, A., Locskai, R., Bohács, G., & Hős, C. (2024). Application of regression models on the prediction of corrosion degradation of a crude oil distillation unit. *Advances in Materials Science*, vol. 24, pp. 74-85. doi:10.2478/adms-2024-0005

Zagorowska, M., Ditlefsen, A-M., Thornhill, N F., & Skourup, C. (2019). Turbomachinery Degradation Monitoring Using Adaptive Trend Analysis. *12th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems*. April 23-26, Florianopolis, Brazil. https://doi.org/10.1016/j.ifacol.2019.06.141

BIOGRAPHIES

Unnat Mankad is a Staff engineer, Mathematics and Data Science, Service Engineering at Baker Hughes, Bengaluru, India. He received his master's degree in mechanical engineering from Birla Institute of Technology and Science (BITS), Pilani. In his current role, Unnat develops complex analytics for on-line prognostics, diagnostics, and predictive maintenance of Baker Hughes' iCenter covered turbo machines. As expert of time series sensor data analysis, he also supports data science and machine learning driven algorithm developments focusing on emissions, reliability, and availability improvements.

Gabriele Mordacci is the iCenter engineering Manager at Baker Hughes, Kuala Lumpur, Malaysia. He received his master's degree in Aerospace engineering from Universita degli studi (UNIPI), Pisa Italy, with a final thesis on the High speed gearbox diagnostic with microphone dynamic analysis. He worked in RMD developing the first analytics based on the fleet statistical for different type of technology and translating the field experience in digital application. As OEM expertise supports diagnostic and data scientist to prepare and validate the new analytics algorithm and output.

Aidil Fazlina Binti Hasbullah *is a Diagnostic Engineer at Baker Hughes, Kuala Lumpur, Malaysia. She received her former tertiary education from Universiti Teknologi PETRONAS (UTP) in Electrical & Electronic Engineering major in Instrumentation and Control. In her current role, Aidil Fazlina focuses on remote monitoring and diagnostics of turbo machines of global installed fleet. She also act as focal point for onshore LNG plants for Malaysia and India fleet. Prior joining Diagnostic team, she has experience of working with LNG plant as Instrument engineer, leading Factory Acceptance test for instrument control system. She also actively collaborates with Baker Hughes' Data Scientists to support advanced analytics development to improve the reliability and availability of the turbo machines*.

Fahzzira Jaafar is a Product Service Engineer at Baker Hughes, Kuala Lumpur, Malaysia. She received her bachelor's degree in mechanical engineering from Universiti Teknologi Malaysia (UTM). Fahzzira has expertise in Balance of plant and a strong experience in monitoring, analysis and troubleshooting of different type of turbomachines covered by Baker Hughes's RMD services. Prior joining Product Services team, she has been involved in number of EPCC (Engineering, Procurement, Construction and Commissioning) and DED (Detailed Engineering Design) projects for Offshore and Onshore applications. In her current role, she is responsible for all activities relating to enhancing services technology integrating customer data, or capturing engines/products reliability, availability, maintenance, safety and other performance parameters.

Carmine Allegorico is a Senior Principal engineer and experienced data scientist at Baker Hughes, Firenze, Italy. He received his master ' s degree in mechanical engineering from University of Napoli Federico II. In his current role, Carmine is a technical point of reference for the analytics discipline providing engineering guidance to other teams, helping to train new engineers and keeping abreast of industry trends and issues. He provides consulting during the development and implementation of advanced solutions for the on-line diagnostic and predictive maintenance, coordinates the creation of internal processes and support the adoption of new platforms and technologies

Gionata Ruggiero is currently covering the position of Asia Pacific Service technology leader and he is currently based in Kuala Lumpur where he lives with his family. In his 20+ year of international experience having worked in Florence, Nigeria, Angola and India and covered several positions in the engineering organization: Subject Matter of expert, Project engineering serving different NOC and IOC, and Monitor & Diagnostic bringing innovative idea with an inclusive approach. Leader of a multicultural and multidisciplinary team, Gionata is responsible of the engineering support since the Installation and commissioning until the end of the life cycle included the IET iCenter in Kuala Lumpur. He is known as Customer focus engineering able to identify solutions and pioneering new technology injection mixing SME and digital domains. Gionata hold the Bachelor of Industrial Engineering and he recently completed advanced studied in Digital transformation and Clime Change Toward Net Zero Emission.

Case Study of Product Development through Generative Design according to Anemometer Replacement Cycles

Joongyu Choi [1], Soyoung Shin [2], and Sangboo Lee [3]

[1,2,3] *RAONX SOLUTIONS INC., Busan & Bundang, South Korea*

jgchoi@raonx.com
syshin@raonx.com
sblee@raonx.com

ABSTRACT

Product Lifecycle Management (PLM) systems are commonly used to manage various product data generated throughout the product lifecycle. This paper explains the results obtained by multiple participants using commercial software within the PLM environment to perform structural and vibration analyses of an Anemometer. Generative design techniques were employed for 3D CAD modeling of the Anemometer, and the commercial analysis software NASTRAN was used for simulation analyses. The open-source PLM system ARAS Innovator's project and workflow management modules were utilized to manage the generated design data, allocate tasks among participants, and control schedules. Through this approach, we propose a method to predict and manage the replacement cycle of Anemometer.

Key Words: Generative Design, PLM, Nastran, ARAS Innovator

1. INTRODUCTION

Currently, the technology for generative AI is very active and progressing at a very fast pace. [1] This direction is also being applied to industrial companies to reflect generative design [2], and this research is being conducted through a research project as described later.

This study is still a work in process, and the model applied in this study evaluated the structural stability of an anemometer, one of the products of the client company, and presented a case of applying it to ARAS Innovator, an OPEN PLM solution, as a management method for a large number of design plans generated through generative design. He is currently conducting research by expanding its application to assembly design in the aero/defense field.

2. DISCUSSION

2.1. Structural Analysis

Fig. 1 shows an anemometer that was damaged during operation. To analyze it, we performed a structural analysis as shown in Fig. 2, a structural analysis was performed.

Fig. 1 Breakage of the Anemometer cup

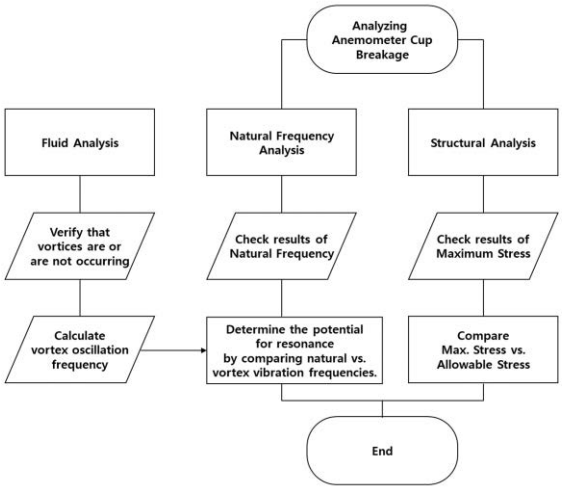


Fig. 2 Analysis process flow chart

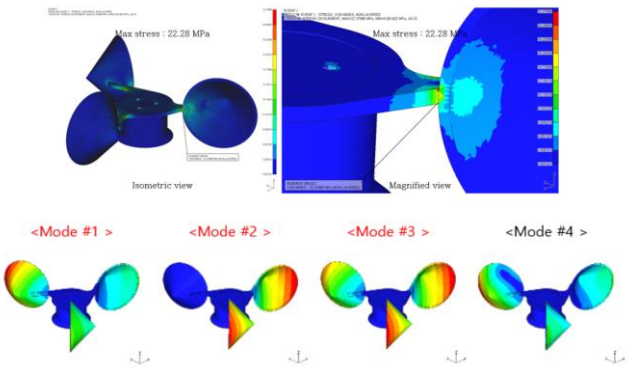


Fig. 3 Structural & Natural Frequency Analysis

The analysis showed that the natural frequency analysis of the structure should be designed to avoid the vortex-induced vibration frequencies around the anemometer cup, and a generative design was derived to satisfy these design criteria.

2.2. Generative Design Draft

Fig. 4 shows a schematic of the design generation process. It shows the process of generating a large number of designs using generative design methods and then optimizing them to find the optimal design. This detailed study is in progress through joint research [3].

Fig. 4 at the bottom, Generative Adversarial Networks with Boundary Constraints (GAN-BC), a deep learning-based model for reverse engineering, is shown. The samples generated from the model and their prediction performance were considered, and supervised learning was performed using the data extracted for prediction and used as a surrogate model. This improved the engineering performance and additive manufacturing suitability of the design created by learning in a direction that minimizes the predicted performance value.

1,000 samples were generated by learning GAN-BC, and compared to existing randomly generated samples, the average weight decreased by about 35.6% from 2.16 kg to 1.39 kg, the average amount of support also decreased by about 21.6%, and the defined It was noticed that the evaluation criteria had improved.

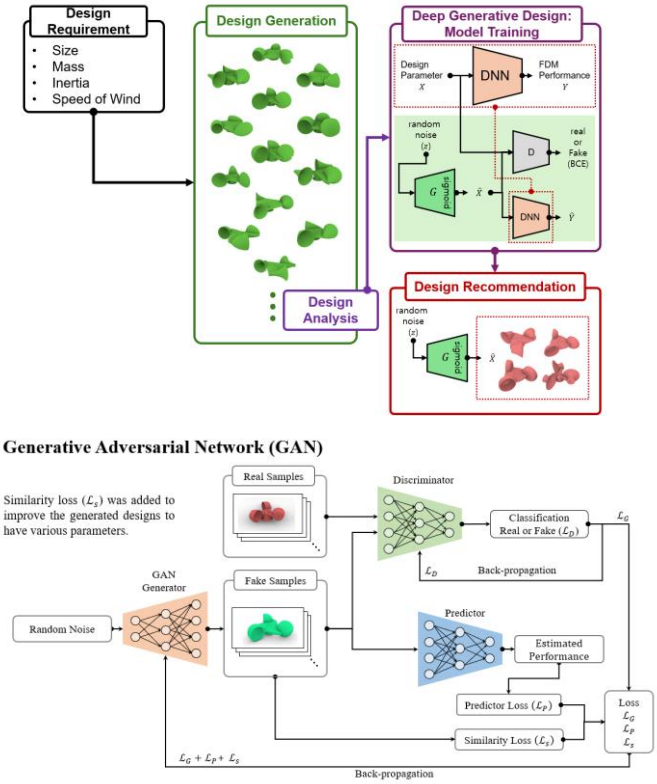


Fig. 4 Generative Design Framework

2.3. Adapting to an open PLM system

The massive number of design alternatives generated through generative design techniques necessitates the establishment and management of a database containing the characteristics of each design alternative. Additionally, procedures such as history management for items reviewed at each step of the workflow will be required.

To implement these elements, we applied the open-source PLM system ARAS INNOVATOR to our research. In Fig. 5, we created a Workflow Manager environment for the design process using the Workflow Map module. Fig. 6 demonstrates an example of establishing a database for the design alternatives.

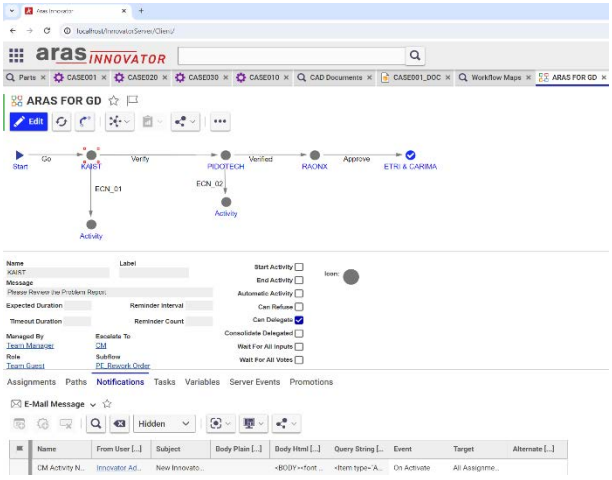


Fig. 5 Workflow Manager for Generative Design

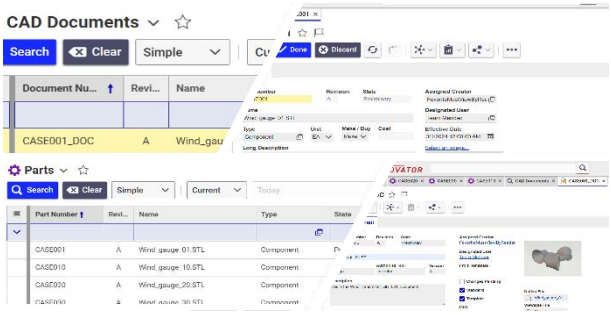


Fig. 6 Open PLM System Application Cases

In the case of workflow manager development, the first year was conducted in an environment where the outputs from each joint organization were integrated, and the second year was conducted to apply it to the PLM system. Currently, research on embedding it in the CAD system is in progress through the third year.

3. CONCLUSION

So far, we have introduced the analytical evaluation and generative design of anemometer, as well as the management plan for multiple design alternatives. Although there were limitations in terms of functionality due to its open-source nature, it is believed that systematic management data for the generated designs will play an important role. Subsequent research is ongoing to apply the final design selected through optimization to production via 3D printing [4]. Based on the technologies developed through such application cases, we are currently conducting generative design for wearable devices in the aerospace and defense industries. Additionally, we are actively promoting our work to secure demand from companies in the electrical and electronics sectors.

4. ACKNOWLEDGMENTS

This paper is based on research funded by the Ministry of Science and ICT and supported by the National Institute of Information and Communication Planning and Evaluation. (No. 2022-0-00969, Development of AI-based generative design technology and production-linked technology for automatically generating large amounts of engineering optimal designs)

REFERENCES

[1] Wang, L., Chan, Y. C., Ahmed, F., Liu, Z., Zhu, P., & Chen, W. (2020). "Deep generative modeling for mechanistic-based learning and design of metamaterial systems." Computer Methods in Applied Mechanics and Engineering, 372, 113377

[2] Kim, S., Jwa, M., Lee, S., Park, S. and Kang, N. (2022). "Deep learning-based inverse design for engineering systems: multidisciplinary design optimization of automotive brakes." Structural and Multidisciplinary Optimization, 65(11), p.323

[3] Jihoon Kim, Sumin Lee, Namwoo Kang (2023). "Deep Learning-based Parametric Inverse Design Considering Mechanical Performance and Additive Manufacturing", Korean Society of Mechanical Engineers.

[4] Eunseo Lee, Changbeom Kim, Hyunchul Kang, Mingi Kim, (2023). "A Study on the Latest Technology Trends and AI Application Method for Automated Additive Manufacturing", Korean Society for Computational Design and Engineering

Feature Selection Method for Gear Health Indicator Using MIC Ranking

Hongliang Song[1], Hongli Gao[2], Ruiyang Zhou[3], Jianing He[4],and Mengfan Chen[5]

[1,2,3,4,5]*Southwest Jiaotong University, Chengdu, Sichuan, 610000, China*

cdsonghl@my.swjtu.edu.cn
hongli_gao@home.swjtu.edu.cn
bk2019111342@my.swjtu.edu.cn
hjn1102140379@163.com
chenmengfan1997@outlook.com

ABSTRACT

In the construction of health indicator for electromechanical equipment, selecting features that exhibit monotonicity, trend characteristics, and a strong correlation with equipment health is paramount to accurately reflect these indices. With the advent of numerous libraries and models for time-series data feature extraction, the range of potential features has expanded significantly. Despite this proliferation, there is a lack of extensive research on effective feature selection. This paper investigates the efficacy of the Maximum Information Coefficient (MIC) method in extracting features that align with the monotonicity and trend-related requirements of electromechanical equipment health indicator. Our experiments indicate that the MIC method adeptly identifies pertinent features for the construction of these indices, underlining its utility in the field of health monitoring for electromechanical systems.

1. INTRODUCTION

The construction of health indicator is essential for evaluating the current health status of engineering systems and their critical components, playing a pivotal role in inferring their Remaining Useful Life (RUL). The accuracy of RUL predictions hinges on the ability to develop health indicator that precisely reflect the condition of these components. Gears, for instance, are key elements in transmission systems. Damage to gears can lead to severe economic losses and potential personnel injuries. Therefore, accurately assessing their health status is crucial to prevent accidents caused by gear failures. This underlines the importance of reliable health indicator construction as a preventative measure against unforeseen mechanical breakdowns.

Supported by Sichuan Science and Technology Program, NO:2023YFG0030
E-mail addresses: hongli_gao@home.swjtu.edu.cn (H. Gao).

Currently, in the fault diagnosis and predictive analysis of gears, the application of vibration signals collected by accelerometers is the most widespread. Compared to other types of signals, such as temperature and pressure, vibration signals exhibit a higher sensitivity in detecting changes in gear health status [1]. The typical process for constructing health indicator, as illustrated in Figure 1, comprises four distinct stages: data acquisition, feature extraction, feature selection, and feature fusion, culminating in the construction of the health indicator. This structured approach ensures a comprehensive analysis, leveraging the sensitivity of vibration signals to accurately reflect the health status of the gears.

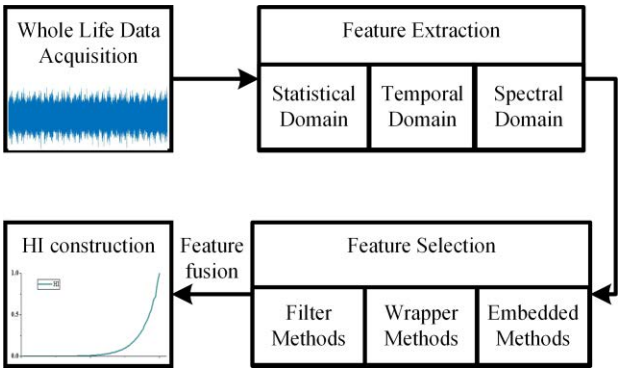


Figure 1. Typical Process for Health Indicator Construction.

In the construction of health indicator, feature extraction methods predominantly yield three types of features [2]-[3]. The first type, statistical domain features, are derived through statistical analysis to capture key characteristics of the data. They describe central tendencies, distribution ranges, and deviations in data shape. The second type, temporal domain features, focus on analyzing changes and dynamic properties in time series data. Finally, spectral domain features are identified through frequency analysis, uncovering periodic

components and spectral distributions within the data. Techniques like Fourier transform and other spectral analysis methods are employed to extract frequency components, which are crucial for understanding oscillatory patterns and frequency-related characteristics in the data. These three feature types compress information carried by the original signal from different perspectives. In health indicator construction, they play pivotal roles, complementing and interrelating with each other to provide a robust feature foundation for a comprehensive assessment of health conditions.

In the context of constructing health indicator, three principal methods are employed for feature selection: filter methods, wrapper methods, and embedded methods [4]. Filter methods involve selecting features based on specific metrics, with the selection process operating independently of the health indicator construction algorithm. This approach prioritizes features based on their statistical properties. In contrast, wrapper methods iteratively utilize the algorithm to assess the impact of different feature sets on the performance of the health indicator. This process iteratively evaluates and selects features based on their contribution to the model's effectiveness. Finally, embedded methods integrate feature selection directly into the algorithm's internal structure. This approach leverages the intrinsic properties of the algorithm to optimize feature selection concurrently with model training, leading to a more cohesive and efficient feature selection process.

Filter methods operate independently of any health indicator construction algorithms. In the context of health indicator construction, filter methods generally rely on a single metric for feature evaluation or employ an average of 2-3 metrics to determine the ranking. Medjaher et al. [5] introduced a novel hybrid feature significance ranking metric in their feature evaluation, incorporating monotonicity, correlation, and robustness for Health Indicator selection. Sun et al. [6] proposed the TWM-U2PL, consisting of a teacher model and a student model. The teacher model includes two independent classifiers that assist in extracting and categorizing wear features. Hu et al. [7] presented a method using minimum Redundancy Maximum Relevance (mRMR) to measure the similarity between features and the correlation between features and categories, facilitating the selection of dimensionless indices. Anil Kumar et al. [8] extracted statistical features from time-domain, frequency-domain, and time-frequency domain signals. They identified important features by calculating feature scores based on the differences in feature values between nearest neighbor pairs of instances.

In the process of constructing health indicator, information theory has been applied to enhance the effectiveness of fault feature extraction and health indicator formulation. Akhand Rai et al. [9] utilized multiscale fuzzy entropy extracted from vibration signals as fault features. These multiscale fuzzy entropy feature vectors form probability distributions. The Jensen-Rényi divergence technique is then applied to differentiate the probability distributions of degraded and healthy multiscale entropy feature vectors, thereby establishing the desired health indicator. Sui et al. [10] proposed a bearing RUL prediction method using Mutual Information (MI) and Support Vector Regression (SVR) models to accurately assess the degradation state of mechanical equipment and comprehend bearing RUL information. Ekhi Zugasti et al. [11] introduced feature selection methodologies using Principal Component Analysis (PCA), Uniform Minimum Redundancy Maximum Relevance (UmRMR), and a combination of both, aimed at resolving the damage detection problem. These approaches demonstrate the value of information-theoretic techniques in creating more accurate and reliable health indicator for mechanical systems.

Selecting features based on criteria such as monotonicity and correlation poses a challenge in effectively gauging the relative importance of each metric. This paper introduces a feature selection method for health indicator utilizing the MIC ranking, which is adept at identifying features that encapsulate a comparatively higher quantity of degradation information. The structure of the remainder of this paper is as follows: Section 2 details the proposed MIC-based health indicator feature selection method. Section 3 describes the experimental setup and data acquisition process. Experimental results are presented in Section 4. Conclusions are drawn in Section 5.

2. METHODOLOGY

In the construction of health indicator, feature selection constitutes a crucial aspect. Given the plethora of feature extraction methods available, it is imperative to selectively identify features that accurately represent the state of degradation. Such features typically necessitate possessing two key attributes: monotonicity and correlation. These attributes are quantifiable and can be effectively measured using specific formulas, designated as Eq. (1) for monotonicity and Eq. (2) for trendiness, as detailed in the referenced literature [12].

Monotonicity primarily measures the trend of a feature, whether it is consistently increasing or decreasing. A feature with the higher monotonicity indicates the better degradation with an increasing/decreasing trend. The calculation of monotonicity is conducted as follows:

$$Mon(f_i) = \left| \frac{\#(\Delta f_i \geq 0)}{L-1} - \frac{\#(\Delta f_i < 0)}{L-1} \right| \quad (1)$$

where $Mon(f_i)$ is the monotonicity value for the i^{th} feature f_i with length of L. $\Delta f_i = f_{i+1} - f_i$ is the difference between consecutive elements. $\#(\Delta f_i \geq 0)$ represents the number of non-negative differences in the f_i sequence. $\#(\Delta f_i < 0)$ represents the number of negative differences in the f_i sequence.

Correlation as a metric primarily reflects the degree of correlation between a feature and the time of degradation. The formula for calculating correlation is as follows:

$$Corr(f_i, T_i) = \left| \frac{cov(f_i, T_i)}{\sigma_{fi} \cdot \sigma_{Ti}} \right| \quad (2)$$

where cov is the covariance of i^{th} feature f_i with the time vector T,and σ is the standard deviation.

An effective understanding of the concepts of monotonicity and correlation in feature analysis can be easily achieved by referring to Figure 2. This figure is divided into two parts: the left side depicts the behaviors of four distinct features, labeled F1 through F4, across their entire lifecycle. The right side, in contrast, illustrates the corresponding Monotonicity Score and Correlation Score for each of these features. By examining these graphical representations, one can clearly discern how different features exhibit varying levels of monotonicity and correlation over time.

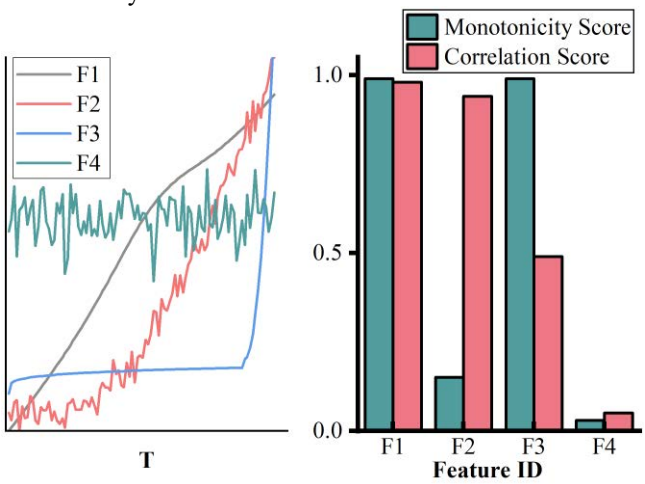


Figure 2. Four representative features. F1 represents high Monotonicity and high Correlation, F2 represents low Monotonicity and high Correlation, F3 represents high Monotonicity and low Correlation, and F4 represents Low Monotonicity and Low Correlation.

The feature selection method for gear health indicators with MIC proposed in this paper is able to complete the feature selection quickly and, at the same time, ensure the monotonicity and trend of the features to a certain extent.

2.1. Basic theory of The MIC

The calculation of the MIC [13] necessitates the computation of mutual information values between variables. Mutual information is a concept in information theory that quantifies the degree of mutual dependence between two random variables. It serves as a measure of the amount of information one variable contains about another. The greater the mutual information value, the stronger the interdependence between the two variables. When considering two random variables, X and Y, their mutual information, denoted as $I(X,Y)$, is defined as follows:

$$I(X,Y) = \sum_{x \in X, y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (3)$$

Where $p(x,y)$ represents the joint probability distribution of X and Y, $p(x)$ and $p(y)$ denote the marginal probability distributions of X and Y.

Unlike mutual information, the MIC demonstrates heightened sensitivity to a broader range of relationship types between variables. It is adept not only at identifying linear and non-linear functional relationships, such as exponential and periodic, but also at detecting non-functional relationships, including combinations or overlays of functional relationships. The aim of MIC is to provide a unified measure of similarity for various types of relationships. MIC builds upon the concept of mutual information. It operates by exploring all possible grid partitions of the data, seeking the partitioning that maximizes the mutual information. The value of MIC ranges between 0, indicating no relationship, and 1, signifying a perfect correlation. This range provides a clear and quantifiable indication of the strength and nature of the relationship between the variables. The MIC functions by calculating mutual information across a range of different grid partitions, with the objective of identifying the partition that maximizes this mutual information. Specifically, for a given dataset, the MIC algorithm evaluates various grid sizes and configurations. It systematically computes the mutual information for each of these configurations. The configuration that yields the highest mutual information is then selected, and its corresponding mutual information value is designated as the MIC value.

In a dataset comprising data points with two attributes, X and Y, these points are distributed within a two-dimensional space. To analyze these data, an $m \times n$ grid is utilized to partition this space. The frequency of data points falling within a specific row x of the grid is used to estimate the marginal probability $p(x)$. Similarly, the frequency of data points in a particular column y is used as an estimate for the marginal probability $p(y)$. Furthermore, the frequency of data points located within a specific cell (x,y) of the grid provides an estimate for the joint probability $p(x,y)$.

$$p(x,y) = \frac{N(x,y)}{\sum_{i=1}^{m} \sum_{j=1}^{n} N(i,j)} \quad (4)$$

By altering the method and arrangement of the grid partitioning, a range of mutual information values can be generated. This variation is crucial in the process of calculating the MIC.

$$MIC(X,Y) = \max_{m*n \leq n^a} \frac{I(X,Y)}{\log_2 \min(m,n)} \quad (5)$$

Where n represents the scale of the data. The value of the constant a can be set based on experience or scale. The condition $m * n \leq n^a$ is to limit the size of the grid for the purpose of dividing regions. Dividing by $\log_2 \min(m,n)$

completes the normalization of data in different dimensions, ensuring that their values fall within the interval [0,1].

2.2. Features Selection in Health indicator Utilizing MIC Ranking

This paper primarily investigates feature extraction and selection from vibration signals. The features extracted in this study are listed in the accompanying table. For detailed explanations of each feature's significance and technical definitions, readers are referred to literature [2], as this paper focuses on the application rather than the detailed descriptions of these features. It is important to note that some features yield multiple output values. In such cases, each distinct output is assigned a unique Feature ID to facilitate clear identification and analysis.

Table 1. Feature List.

ID	Statistical Domain Features	ID	Temporal Domain Features	ID	Spectral Domain Features
1	Absolute energy	2	Area under the curve	9	Fundamental frequency
4	Average power	3	Autocorrelation	23	Max power spectrum
6-7	ECDF Percentile	5	Centroid	33	Median frequency
8	Entropy	24	Maximum frequency	39	Power bandwidth
10-19	Histogram	27	Mean absolute diff	43	Spectral centroid
20	Interquartile range	28	Mean diff	44	Spectral decrease
21	Kurtosis	31	Median absolute diff	45	Spectral distance
22	Max	32	Median diff	46	Spectral entropy
25	Mean	35	Negative turning points	47	Spectral kurtosis
26	Mean absolute deviation	36	Neighbourhood peaks	48	Spectral positive turning points
29	Median	38	Positive turning points	49	Spectral skewness
30	Median absolute deviation	41	Signal distance	50	Spectral slope
34	Min	54	Sum absolute diff	51	Spectral spread
37	Peak to peak distance	56	Zero crossing rate	52	Spectral variation
40	Root mean square				
42	Skewness				
53	Standard deviation				
55	Variance				

In the context of feature selection for health indicators, it is necessary to first construct a progressively growing sample sequence $\boldsymbol{T} = [1, 2, \ldots, N]$ based on the sampling interval. The feature set composed of features in TABLE I is denoted as $\boldsymbol{F} = \{\boldsymbol{F_1}, \boldsymbol{F_2}, \ldots, \boldsymbol{F_L}\}$. The pseudocode for feature selection is as follows:

Table 2. Based on MIC Health Indicator Feature Selection.

Input: $\boldsymbol{T}, \boldsymbol{F}$
output: $\boldsymbol{F}$ *are sorted by MIC*

1: for each feature $\boldsymbol{F_i} \in \boldsymbol{F}$ do
2: MIC of $\boldsymbol{F_i} = \boldsymbol{0}$
2: for (m, n) such that $m * n \leq n^a$ do
3: Divide $\boldsymbol{T}, \boldsymbol{F_i}$ according to m, n to form a grid G
4: Calculate the mutual information $I(\boldsymbol{F_i}, \boldsymbol{T})$ of $\boldsymbol{F_i}$ and $\boldsymbol{T}$ on grid G
5: Normalized mutual information
6: if Normalized mutual information> MIC of $\boldsymbol{F_i}$
7: MIC of $\boldsymbol{F_i}$ = Normalized mutual information
7: Add MIC of $\boldsymbol{F_i}$ in MIC list
8: Sort $\boldsymbol{F_i}$ in $\boldsymbol{F}$ by MIC list

The two principal characteristics of the MIC offer significant advantages in the context of feature selection for health indicator.

Generality: The MIC demonstrates a high degree of applicability across a wide array of relationship types, encompassing linear, non-linear, monotonic, and non-monotonic associations. This Generality enables the selection of features that are representative of diverse functional relationships, thereby facilitating more effective feature fusion in reflecting health indicator.

Equitability: MIC exhibits a relatively consistent sensitivity across different types of relationships. This means that whether the relationship between variables is linear, curvilinear, or follows other complex patterns, MIC can identify it with similar efficacy, provided the relationship is sufficiently strong. Consequently, MIC is capable of selecting features that are most relevant and informative, enhancing the accuracy and reliability of the resulting health indicator.

3. EXPERIMENTAL PROCEDURE

In this study, the experimental data set was collected through accelerated degradation tests conducted on gears. The experimental platform consisted of a two-stage parallel-axis gearbox. The torque applied to the gearbox was generated by a load motor attached to the output end. An accelerometer was mounted at the output cover to capture vibration signals along the Z-axis of the gearbox. The data collection was conducted with a high sampling frequency of 12,800 Hz, ensuring detailed capture of the vibration characteristics. The input frequency to the gearbox was set at 40 Hz. The platform for accelerated degradation tests conducted on gears is shown in Figure 3. The position relationship of each transmission gear in the gearbox is shown in Figure 4.

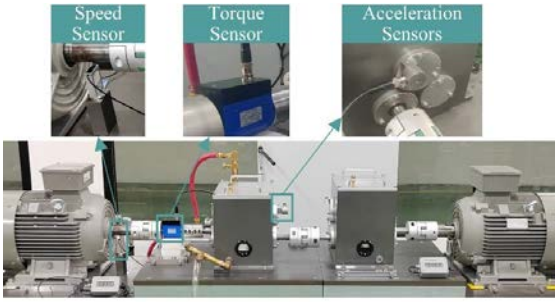


Figure 3. Experimental Platform.

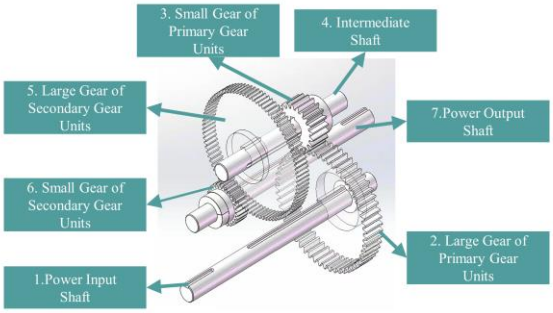


Figure 4. Position Relationship of Each Transmission Gear in the Gearbox

For detailed specifications of the basic gear parameters, readers are directed to Table 3. Additionally, the data set encompasses real-life operational data of gearboxes throughout their entire lifecycle, recorded under three different load conditions. For a more comprehensive understanding of these data sets, including the specific conditions and parameters, please refer to Table 4.

Table 3. Gearbox Parameters.

Parameters	Primary Gear Units	Secondary Gear Units
Number of small gear teeth	29	36
Number of large gear teeth	95	90
Pinion tooth width/mm	15	15
Large gear tooth width/mm	15	15
Modulus/mm	1.5	1.5
Pressure angle/°	20	20

Table 4. Experimental Data Set.

Gear ID	torque load	Total Working Hours (H)	Sample Size
G_1	50%	110	3303
G_2	60%	102	3079
G_3	70%	34	1022

Figure 5 provides a graphical representation of the vibration signals from the tested gearbox, labeled G_1, over its entire lifecycle. The temporal progression of these signals is distinctly illustrated, with noticeable variations becoming evident as time progresses. This variation in the vibration signals is indicative of changes in the gearbox's condition, suggesting a correlation with the performance degradation of the gear.

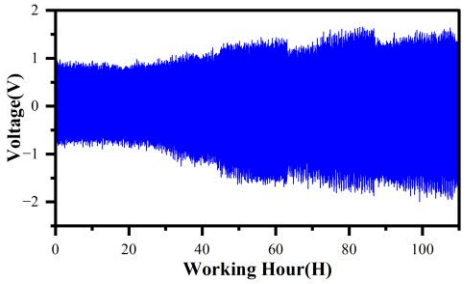


Figure 5. The Z-axis Vibration Signals Over Entire Lifecycle of G_1.

4. RESULT AND DISCUSSION

The study involved extracting a comprehensive set of 56 features from the full lifecycle experimental data of three distinct gear sets. Following the extraction, the feature selection process, as detailed in Section 2.2 of this document, utilized the MIC algorithm. This algorithm was applied to each feature to calculate its MIC value, assessing the strength of the relationship between the feature and the gear's health status. Subsequently, the features were sorted based on the average MIC values computed across the three gear sets, providing a comparative view of their significance. The results of this feature selection and sorting process are illustrated in the Figure 6, offering insights into the relative importance of the various features in the context of gear health monitoring.

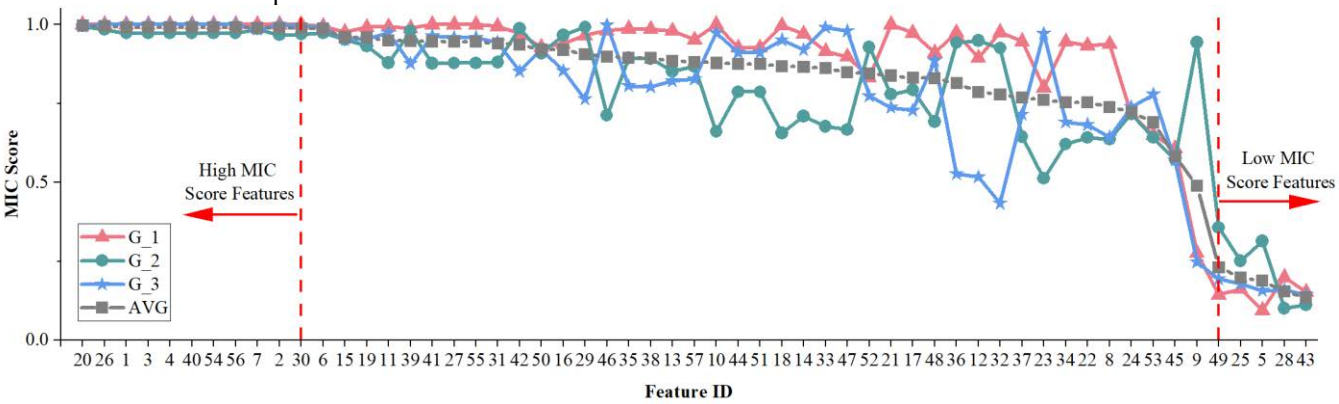


Figure 6. MIC Values for 56 Features of Three Gears.

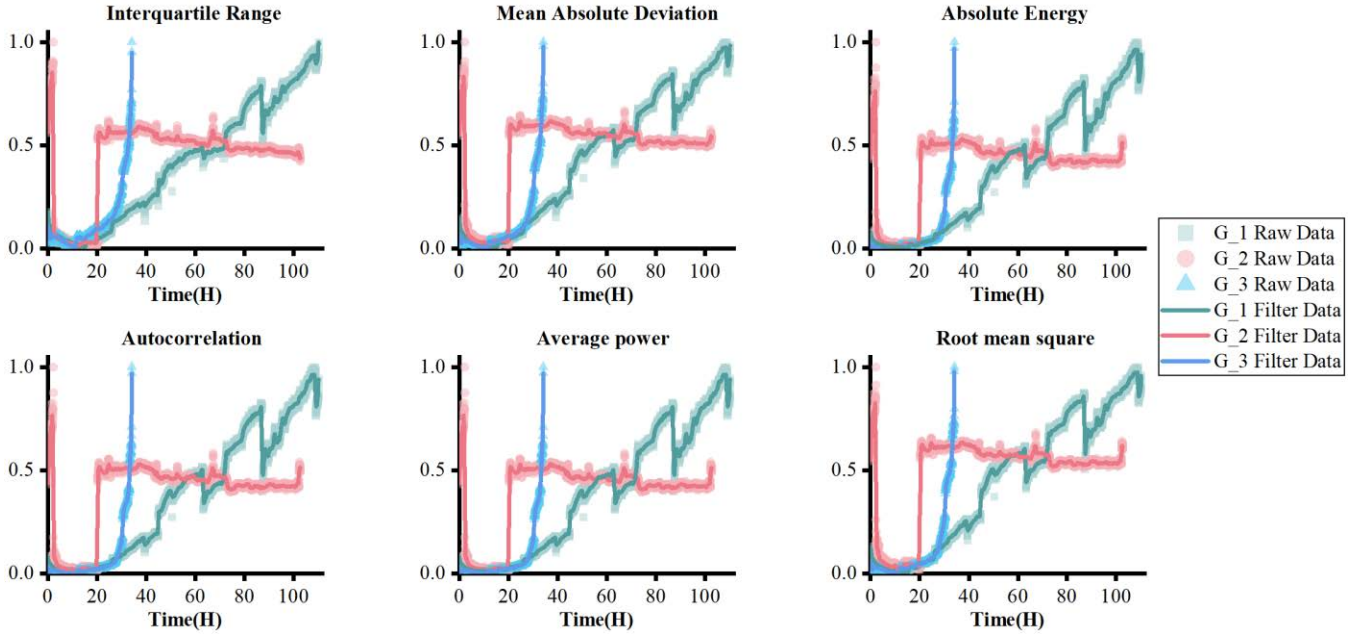


Figure 7. Lifecycle Curves of the Top Six Features Ranked by MIC Score.

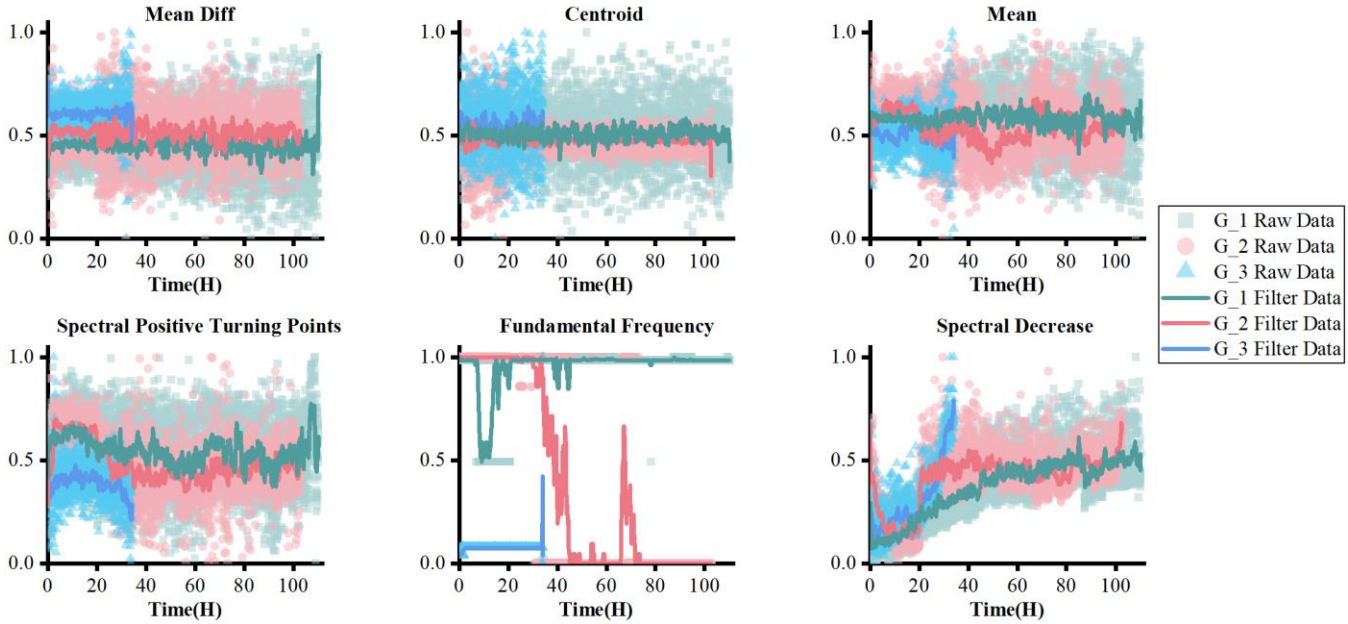


Figure 8. Lifecycle Curves of Features Ranked in the Bottom Six by MIC Score

It is observed that some features exhibit MIC Scores nearing 1 in Figure 6, indicating a significant non-linear relationship between these features and the equipment's degree of degradation. To analyze this further, the features with the top six and bottom six MIC Scores were normalized and their lifecycle variation curves were plotted in Figure 7 and Figure 8. The analysis reveals that the features ranked in the top six display pronounced trendiness and a certain degree of monotonicity, suggesting a strong correlation with the equipment's degradation process. Conversely, the features ranked in the bottom six show little to no discernible trend or pattern. This contrast underscores the efficacy of MIC Scores in distinguishing features that are strongly indicative of equipment health from those that are less informative.

Utilizing Eq. (1) and (2), the monotonicity and trendiness indices of the features were calculated. The analysis revealed a discernible positive correlation between the MIC values and these indices in Figure 9. Specifically, it was observed that features with higher MIC values tend to exhibit more pronounced monotonicity and trendiness. Conversely,

features with lower MIC values generally show weaker performance in these aspects. This correlation indicates that the MIC can be a reliable indicator of a feature's relevance, particularly in terms of its monotonic and trend-based behavior, which are critical attributes in assessing the health and degradation of equipment.

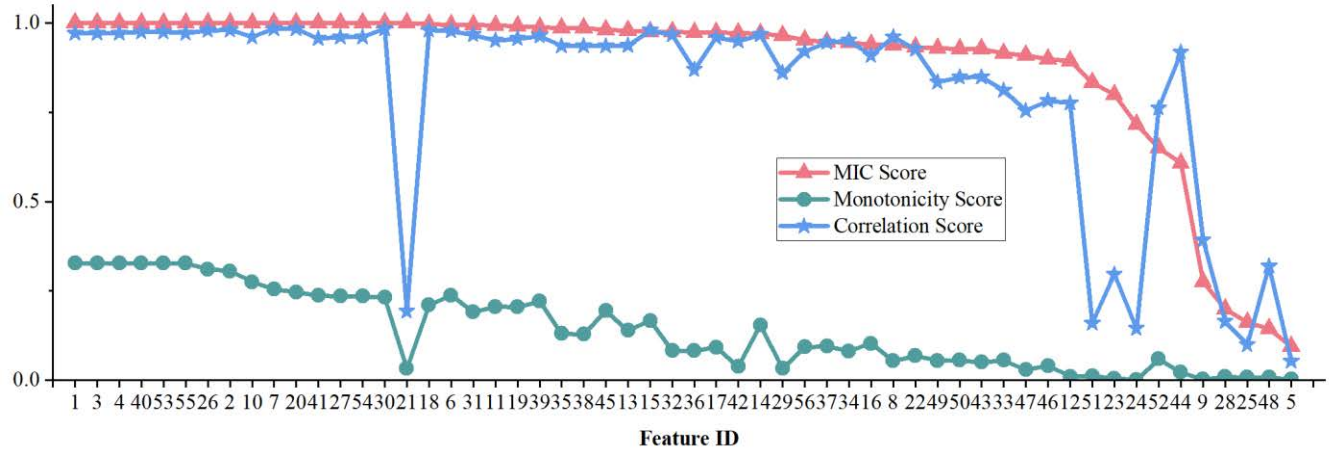


Figure 9. MIC Score, Monotonicity Score and Correlation Score.

5. CONCLUSION

The application of the MIC algorithm in this study has proven to be highly effective in selecting features that correlate closely with gear health. This approach ensures that the resultant health indicator exhibit enhanced monotonicity and trendiness, thereby providing a more accurate reflection of the gear's condition. Notably, MIC also effectively compensates for the shortcomings of mutual information by offering a more comprehensive quantification of the correlation between features and equipment health.

However, it is important to acknowledge a key limitation of the MIC algorithm: its reliance on large datasets for meaningful computation. The efficacy of MIC is significantly reduced when applied to smaller datasets. Recognizing this constraint, future research efforts will focus on modifying and improving the algorithm to better suit applications involving smaller data samples. Such advancements will broaden the applicability of this method, allowing for more versatile and reliable gear health assessments across a wider range of data scenarios.

REFERENCES

[1] D. Wang, K.-L. Tsui, and Q. Miao, "Prognostics and Health Management: A Review of Vibration Based Bearing and Gear Health Indicators," *IEEE Access*, vol. 6, pp. 665–676, 2018, doi: 10.1109/ACCESS.2017.2774261.

[2] M. Barandas *et al.*, "TSFEL: Time Series Feature Extraction Library," *SoftwareX*, vol. 11, p. 100456, Jan. 2020, doi: 10.1016/j.softx.2020.100456.

[3] Y. Sun *et al.*, "Transfer learning: A new aerodynamic force identification network based on adaptive EMD and soft thresholding in hypersonic wind tunnel," *Chinese Journal of Aeronautics*, vol. 36, no. 8, pp. 351–365, Aug. 2023, doi: 10.1016/j.cja.2023.03.024.

[4] J. Li *et al.*, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, p. 94:1-94:45, Dec. 2017, doi: 10.1145/3136625.

[5] V. Atamuradov, K. Medjaher, F. Camci, N. Zerhouni, P. Dersin, and B. Lamoureux, "Machine Health Indicator Construction Framework for Failure Diagnostics and Prognostics," *J Sign Process Syst*, vol. 92, no. 6, pp. 591–609, Jun. 2020, doi: 10.1007/s11265-019-01491-4.

[6] Y. Sun, H. Gao, J. He, L. Guo, and H. Song, "A New Semi-Supervised Tool-Wear Monitoring Method Using Unreliable Pseudo-Labels." Rochester, NY, Sep. 07, 2023. doi: 10.2139/ssrn.4564476.

[7] Q. Hu, X.-S. Si, A.-S. Qin, Y.-R. Lv, and Q.-H. Zhang, "Machinery Fault Diagnosis Scheme Using Redefined Dimensionless Indicators and mRMR Feature Selection," *IEEE Access*, vol. 8, pp. 40313–40326, 2020, doi: 10.1109/ACCESS.2020.2976832.

[8] A. Kumar *et al.*, "A novel health indicator developed using filter-based feature selection algorithm for the identification of rotor defects," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 236, no. 4, pp. 529–541, Aug. 2022, doi: 10.1177/1748006X20916953.

[9] A. Rai and J.-M. Kim, "A Novel Health Indicator Based on Information Theory Features for Assessing Rotating Machinery Performance Degradation," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6982–6994, Sep. 2020, doi: 10.1109/TIM.2020.2978966.

[10] W. Sui, D. Zhang, X. Qiu, W. Zhang, and L. Yuan, "Prediction of Bearing Remaining Useful Life based on Mutual Information and Support Vector Regression Model," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 533, no.

1, p. 012032, May 2019, doi: 10.1088/1757-899X/533/1/012032.
[11] E. Zugasti, L. E. Mujica, J. Anduaga, and F. Martínez, "Feature Selection - Extraction Methods Based on PCA and Mutual Information to Improve Damage Detection Problem in Offshore Wind Turbines," *KEM*, vol. 569–570, pp. 620–627, Jul. 2013, doi: 10.4028/www.scientific.net/KEM.569-570.620.
[12] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, May 2017, doi: 10.1016/j.neucom.2017.02.045.
[13] D. N. Reshef *et al.*, "Detecting Novel Associations in Large Data Sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011, doi: 10.1126/science.1205438.

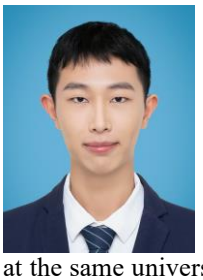
BIOGRAPHIES

Hongliang Song obtained M.S. degree in Mechanical and Electronic Engineering from Southwest Jiaotong University, located in Chengdu, Sichuan, China, in 2020. He is currently pursuing Ph.D. in the same department at Southwest Jiaotong University. Throughout his academic journey, he has been deeply involved in research related to the fault diagnosis and intelligent maintenance technology of electromechanical equipment.

Hongli Gao received the Ph. D degree in mechanical engineering from Southwest Jiaotong University, Sichuan, P. R. China in 2005.

He is currently a Professor at Southwest Jiaotong University. His research interests focus on designing and the reliability analysis of complex electromechanical equipment.

Zhou Ruiyang obtained B.S. degree in Mechanical Design, Manufacturing and Automation from Southwest Jiaotong University, located in Chengdu, Sichuan, China, in 2023. He is currently pursuing M.S. degree in the Department of Mechanical and Electronic Engineering at Southwest Jiaotong University. He is actively involved in research related to the fault diagnosis and intelligent maintenance technology of electromechanical equipment.



Jianing He earned his B.S. degree in Mechanical Design, Manufacturing and Automation from Southwest Jiaotong University, situated in Chengdu, Sichuan, China in 2023. He is now advancing his studies by pursuing an M.S. degree in the Department of Mechanical and Electronic Engineering at the same university. His main area of research is centered around the fault diagnosis and intelligent maintenance technology of electromechanical equipment.

Mengfan Chen obtained a Bachelor's degree in Industrial Engineering from the School of Mechanical Engineering at Southwest Jiaotong University in Chengdu, Sichuan, China, in 2019. Throughout his academic journey, he has been deeply involved in research related to the intelligent monitoring and diagnosis of equipment states, as well as the development of industrial software testing.

Filter-based feature selection for prognostics incorporating cross correlations and failure thresholds

Alexander Löwen[1], Peter Wissbrock[2], Amelie Bender[1], and Walter Sextro[1]

[1] *Paderborn University, Faculty of Mechanical Engineering, Chair of Dynamics and Mechatronics, Paderborn, 33098, Germany*
alexander.loewen@uni-paderborn.de
amelie.bender@uni-paderborn.de
walter.sextro@uni-paderborn.de

[2] *Lenze SE, Innovation Department, Aerzen, 31855, Germany*
peter.wissbrock@lenze.com

ABSTRACT

Historical condition monitoring data from technical systems can be utilized to develop data-driven models for predicting the remaining useful life (RUL) of similar systems, whereas the Health Index (HI) often is a crucial component. The development of robust and accurate models requires meaningful features that reflect the system's degradation process, enabling an accurate prediction of the system's HI. Traditionally, the identification of those is supported by one of various feature ranking methods. In literature, feature interdependencies and their transferability across various similar systems are not sufficiently considered in feature selection, exacerbating the challenge of HI prediction posed by the scarcity of data and system diversity in real-world applications. This work addresses this gaps by demonstrating how filter-based feature selection, incorporating failure thresholds and cross correlations, enhances feature selection leading to improved HI prediction. The proposed methodology is applied to a novel dataset* obtained from run-to-failure experiments on geared motors conducted as part of this study, which presents the aforementioned challenges. It is revealed that classical feature selection, consisting of feature ranking only, leaves potential untapped, which is utilized by the proposed selection methodology. It is shown that the proposed feature selection methodology leads to the best result with a RMSE of 0.14 in predicting the HI of a constructive different gearbox, while the features, determined by classical feature selection, lead to a RMSE of 0.19 at best.

* The dataset called Lenze-GD is accessible via: https://doi.org/10.5281/zenodo.11162448.

1. INTRODUCTION

Early fault detection of mechanical systems like gears and motors is an important topic for industrial production, helping companies to predict equipment failures, reduce downtime and to ensure the reliability and safety of industrial systems. The analysis of data from time series sensors like acoustic, vibration, position, or current is of great interest to monitor the health condition of machines and to predict failure in the mechanical systems life-cycle. The prognostics of the remaining useful lifetime (RUL) aims to predict operating time of a typical operational lifespan that a mechanical system has already passed and estimate the amount of the remaining useful life. In particular, vibration signals have been widely used for RUL-prognostics. However, the usage of signals acquired from inverts like the motor current reduces costs of installation and maintaining external sensors. Under the limitation of a drive system including an induction motor and an inverter with a sufficient data interface, the motor becomes the sensor.

A major challenge in developing accurate and robust RUL-prognostics is the limitation of data, especially in scenarios where abnormal observations are rare or difficult to obtain, referred to as data scarcity. In this study geared motors are focused, which are combinations of toothed-wheel-based gearboxes and of electric induction motors. To match the diverse requirements of customers, the geared motors can be configured and scaled individually. These customized geared motors can be used in a variety of different machine types, which also may be customized. In many real-world problems it is realistic that only a few or none run-to-failure data-collections are available and thus often only data from the healthy motor can be used for model training.

The work is structured as follows. Section 2 presents a comprehensive feature engineering methodology with focus on

feature selection to overcome data scarcity and address system differences. A multi-stage feature selection methodology is described followed by the machine learning (ML) models used and trained based on the selected features. ML algorithms employed are Gaussian Process Regression (GP), Linear Regression (LR), Multi-Layer-Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM). Next, a novel dataset from run-to-failure experiments on geared motors including gear-mesh and bearing failures is introduced in section 3. The experimental setup and the recorded data, which are obtained from a frequency inverter, are described. In section 4, the proposed feature engineering methodology is applied on the new data. In section 5 the advantages over the classical feature selection, consisting of feature ranking only, is shown resulting in the best root mean squared error (RMSE) of 0.14, in contrast to the classical selection's best RMSE of 0.19 in predicting the health index (HI).

2. METHODOLOGY

In this paper, a broadly used workflow for diagnostics and prognostics of technical systems is utilized, which comprises the elements data preprocessing, feature extraction and diagnostics or prognostics algorithm (Goyal, Mongia, & Sehgal, 2021; Ly, Tom, Byington, Patrick, & Vachtsevanos, 2009). Depending on the application, these elements are generally adapted and optimized to suit the circumstances of any given application. The methodology employed prioritizes a more generalized process. To address this limitation, feature engineering is focused wherein a wide range of features are computed, adapted and a multi-stage feature selection process is adopted to select subsequently the most relevant features. Data-driven algorithms are then trained with the selected features within a cross-validation process that includes hyperparameter optimization to predict the HI of the system. These steps are parameterized by means of the systems used for training and then applied to the system used for testing. The whole process is shown in Fig. 1 comprising feature extraction, feature processing, feature selection and model training including hyperparameter optimization, with particular focus on feature selection, whereas Fig. 2 shows the steps from feature processing to correlation analysis with more detail. The steps are described in the following.

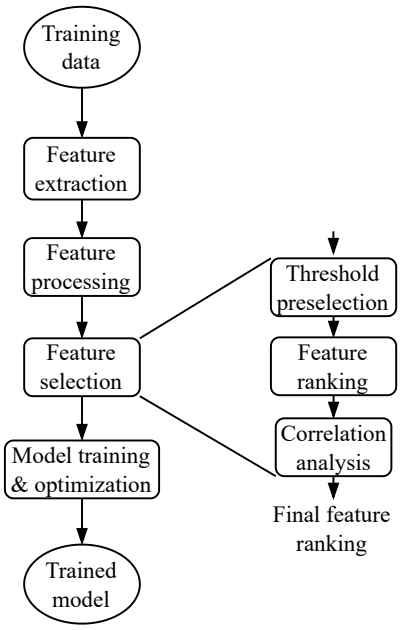


Figure 1. Overview of the applied training process.

2.1. Feature Extraction

Feature extraction is applied to each measurement and channel to extract information regarding system's degradation over time. To address a variety of a system's characteristics, a multitude of features are computed, aiming to encompass a wide range of potential applications where any given feature may capture the system's degradation process. To extract features from time series data, the publicly available Python package tsfresh is used (Christ, Braun, Neuffer, & Kempa-Liehr, 2018). tsfresh is utilized for an automatic extraction of time series features which comprises features from the time, frequency, and time-frequency domains.

In the review of (Goyal et al., 2021), several use cases regarding rotating mechanical systems are consolidated, highlighting the frequent utilization of the fast Fourier transform (FFT) for analysis. This underlines the capacity of FFT to extract information from data, particularly from systems with rotating components, to infer health-related insights. Therefore, additional frequency-dependent features are calculated by dividing the frequency spectrum into sections defined by a constant percentage bandwidth (CPB). Maximum and average FFT coefficients are extracted from the corresponding sections to capture amplitude changes in smaller frequency spectrums. A CPB analysis has been utilized, among other fields, in the field of acoustics (Gram-Hansen, 1991), providing the opportunity to efficiently consider the entire frequency spectrum within comprehensive feature extraction.

2.2. Feature Processing

Feature processing encompasses feature smoothing and feature scaling. Feature smoothing is utilized to reduce noise and variability from the feature data, making underlying patterns and trends more apparent. The moving average is often applied for this purpose. Feature scaling involves scaling the computed features based on the median value of their initial feature data points as shown in Eq. (1). Here, $f_{i,j}$ represents feature i of system j, $f_{i,j,init}$ contains the initial feature points, and $f^*_{i,j}$ denotes the scaled feature data. This process aims to eliminate unwanted influences and facilitate bet-

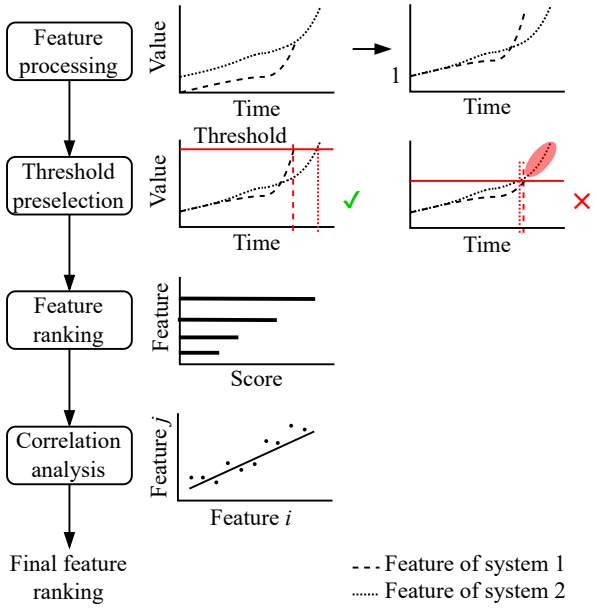


Figure 2. Overview and illustration of the feature processing and selection process.

ter comparability among different systems.

$$f^*_{i,j}(t) = \frac{f_{i,j}(t)}{\text{median}(f_{i,j,init})} \tag{1}$$

2.3. Feature Selection

Feature selection can be divided into filter, wrapper, embedded and hybrid methods (Hoque, Bhattacharyya, & Kalita, 2014) and is necessary for information concentration. This paper focuses on filter methods, as they are less computational intensive in general (Hoque et al., 2014). Feature selection utilized comprises the steps threshold preselection, feature ranking and correlation analysis. Threshold-based preselection retains features with similar failure thresholds and discards those without. It is followed by feature ranking and correlation analyses to remove highly correlated features. The steps are described in the following.

Threshold preselection: Subsequently on feature processing, a preliminary selection is conducted based on a common threshold value for each feature across the systems. Thresholds for system failure are often determined using a predefined HI, often as linear, e.g. in (Yang et al., 2016), or constructed based on selected features, e.g. in (Thoppil, Vasu, & Rao, 2021). In more rare cases, only one feature is directly used if it is sufficient to reflect the degradation process, provided that it can be used to define a system-wide failure threshold, e.g. in (Li, Huang, Gao, Zhao, & Li, 2023; Bender & Sextro, 2021). With limited data, it is difficult to evaluate

Table 1. Metrics considered to determine feature ranking.

Source	Mon.	Trend.	Rob.	Name
(Carino et al., 2015)	×			Spearman
(Nie et al., 2022)	×	×		Cori-Score
(Chen et al., 2019)	×	×	×	MTR_C
(Zhang et al., 2016)	×	×	×	MTR_Z

an individual feature for use as a reliable HI, as well as to construct a HI with respect to a failure threshold. Specific, multiple features that indicate a common threshold across the systems are often not explicitly sought out. To do this, a criterion based on the thresholds for each feature and system is introduced. Firstly, a threshold τ_i regarding Eq. (2) is calculated for each feature i with respect to the systems denoted by j. Here, $\widetilde{f}^*_{i,j,end}$ denotes the median value of the scaled feature points within the final portion, defined by α, of the RUL. The thresholds are reached at different points in the lifetime of each system. If the minimum reached lifetime, as determined by the specified threshold, is below β of the total lifetime of one of the systems, the feature is discarded. An example is given in Fig. 2, where a feature is marked with a cross, signifying that the systems 2 reaches the threshold prematurely, leading to the exclusion of this specific feature.

$$\tau_i = \begin{cases} \min\limits_j(\widetilde{f}^*_{i,j,\text{end}}) & \text{if } \widetilde{f}^*_{i,j,\text{end}} \geq 1 \\ \max\limits_j(\widetilde{f}^*_{i,j,\text{end}}) & \text{if } \widetilde{f}^*_{i,j,\text{end}} < 1 \end{cases} \tag{2}$$

Feature ranking: Feature ranking is crucial in predictive analysis as it allows to identify the most relevant and informative features. Evaluation metrics employed typically encompass assessment of monotonicity and trendability analysis (Carino, Zurita, Delgado, Ortega, & Romero-Troncoso, 2015; Nie, Zhang, Xu, Cai, & Yang, 2022). Moreover, these metrics can be combined with a metric to consider the robustness (Chen, Xu, Wang, & Li, 2019; Zhang, Zhang, & Xu, 2016). A short overview of considered metrics by source to perform feature ranking is given in Tab. 1 and described in the following.

In (Carino et al., 2015) the monotonicity is calculated using the Spearman correlation coefficient, while monotonicity in (Nie et al., 2022; Chen et al., 2019; Zhang et al., 2016) is assessed through the counts of positive and negative derivatives. Trendability is assessed usually through calculating the Pearson correlation coefficient (Chen et al., 2019; Nie et al., 2022; Zhang et al., 2016). Here, (Zhang et al., 2016) used smoothed feature values to encompass monotonicity and trendability, while all others evaluate the original feature data set. The robustness of a feature is assessed through comparison the raw feature values with their smoothed values (Chen et al., 2019; Zhang et al., 2016). The evaluation across multiple considered metrics is conducted using either the average score or the equally weighted sum. In this paper, all of the named fea-

Table 2. Fictional correlation matrix of the best 3 ranked features.

Feature	1	2
1	-	-
2	0.955	-
3	0.892	0.851

ture ranking methodologies are considered to get insight into the potential of the proposed feature selection methodology.

Correlation analysis: Correlation analyses, specifically the Pearson correlation, are often used, besides for feature selection, for similarity analyses (Guo, Li, Jia, Lei, & Lin, 2017; Nie et al., 2022). In this paper, the Pearson correlation is used to determine the similarity between features. Based on the similarity, highly similar features are classified as redundant and discarded, while the best-ranked features are retained. Tab. 2 provides a fictional example showing a correlation matrix for the ranked features 1, 2 and 3. Feature 2 correlates with a coefficient of 0.955 with feature 1. Feature 3 shows correlations of 0.892 and 0.851 with feature 1 and 2 respectively. A parameter can be used to specify which correlation is acceptable. Features that exceed this parameter across all systems are discarded, ensuring that only unique and informative features are retained. If the parameter in the example shown is set to 0.95, feature 2 is discarded, as it exceeds the parameter for feature 1.

2.4. Model training and test

Different ML algorithms are applied and optimized with regard to their hyperparameters applying a Bayesian optimizing algorithm provided by the scikit-optimize library (Head, Kumar, Nahrstaedt, Louppe, & Shcherbatyi, 2021). This technique is based on probabilistic modeling to explore the hyperparameter landscape and find the best parameter combinations (Garnett, 2023). The main goal is not to compare ML algorithms against each other. Instead, the focus lies on assessing the effectiveness and obtaining the best possible prediction result based on the introduced feature selection method. The ML algorithms employed include GP, LR, MLP with one hidden layer consisting of 100 neurons, RF and SVM from the sklearn library (Pedregosa et al., 2011). These algorithms are trained on processed and selected features and are optimized within a cross-validation process to predict a linear HI. The hyperparameter ranges used for optimization are given in the appendix, with standard values employed if a hyperparameter is unspecified.

The predictions are constrained between 0 and 1, where 0 denotes system failure. Evaluation of the models is based on the RMSE as calculated in Eq. (3). Here, $y_{true,i}$ denotes the true and $y_{predicted,i}$ the predicted HI for each observation i of n total observations.

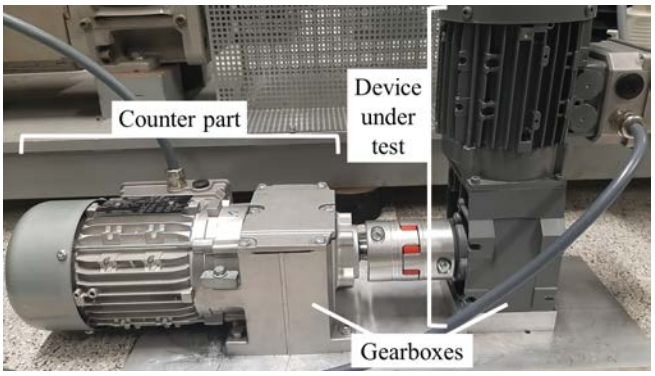


Figure 3. Experimental setup. On the left side, the counter part is shown, which is a helical geared motor. The geared motor on the right side is the device under test, which is a bevel gear.

Table 3. Overview of the gearboxes and their nominal values.

Name	**Type**	**Usage**	**Torque**	**Gear ratio**
H110	Helical	Counter part	110 Nm	28,738
B45	Bevel	Device under test	45 Nm	25,051
H45	Helical	Device under test	35 Nm	10,033

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{true,i} - (y_{predicted,i})^2} \qquad (3)$$

3. Case-Study

To facilitate the presented studies, a run-to-failure experiment for geared motors is introduced. A geared motor is installed in healthy condition and operated until it fails. Throughout the experiment, a data acquisition system is active to monitor the signals of all degradation states. In order to complete the experiment in limited time, the geared motors nominal torque is exceeded. The experiment is conducted three times in total and each with multiple operation states during measurement.

3.1. Experimental Setup

The mechanical part of the setup consist of a first geared motor, the device under test, and a second geared motor, the counter part, shown in Fig. 3. All gearboxes consist of two gear stages with in sum four toothed wheels. The function of the counter part is to create a load for the device under test. An overview of the nominal values of the gears is given in Tab. 3. Thereby the counter part has significant higher nominal torque, to make sure, that the device under test will cause failure, while the counter part stays in healthy condition. The actual torque is selected to lie in the mid of the finite life fatigue of the Woehler characteristic of the second and last gear stage of the device under test to accelerate degradation. During the experiment, the device under test runs with nominal speed.

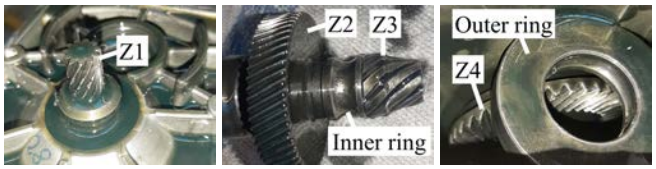


Figure 4. Failure of gearbox B45. Shown are the gears from left to right Z1 with moderate wear, Z2 with minor wear, Z3 with destructive wear and Z4 with moderate wear. As well as the destroyed bearings' inner ring next to Z3 and outer ring next to Z4.

Figure 5. Failure of gearbox H45. Shown are the gears from left to right Z1 and Z2 with minor wear, Z3 with moderate wear and Z4 along with the gearbox full of the deteriorated oil. As well as the destroyed bearings inner ring next to Z3.

In sum, three run-to-failure experiments were conducted, one with B45 gearbox and two with H45 gearbox. In the following, one of the H45 gearboxes will be referred to as H45I and the other as H45II, if they are considered separately. A run-to-failure experiment ends when the gearbox failed, which means its transmission is interrupted. Here, gearbox failure occurred after around 200 hours (H45II) to 790 hours (B45). Subsequently, the gearboxes are opened to evaluate the failures. In the following, the gears are named beginning from the motor-shaft with Z1 transmitting over Z2 to the middle-shaft with Z3 transmitting to the output-shaft over Z4.

The B45 gearbox shows a destructive wear at the Z3, while Z1 and Z4 also show moderate wear, but they stay functional. Z2 only shows minor wear. All of which is shown in Fig. 4. This observation can be explained by the higher torque transmitted by Z3 and Z4 than the first gear stage with Z1 and Z2 and the higher rotation speed of Z1 and Z3 resulting in sum to the high wear of Z3. In addition, the bearing of the middle shaft most close to Z3 is destroyed.

The failure of the runs with H45 shows only minor wear at the gears, except Z3 which shows moderate wear, see Fig. 5. The failures are caused by destroyed bearings next to Z3. Overall, both gear wearing and a destroyed bearing in all cases is observed.

3.2. Data Acquisition

Once per hour, the steady operation of the experiment is interrupted to gather signal measurement of four operation states. These four states are aligned with the nominal values of the induction motor for star connection of the device under test. The states are the combinations of positive or negative nominal speed with nominal or idle torque. Each observation takes measurement with constant sampling rate of 8 kHz and for 2^{15} sampling points, which defines a time period of about 4 s. During this time period, 8 channels are stored in parallel, which are shown in Tab. 4.

Table 4. Overview of the channels acquired by the inverter.

Channel	Type
1	Direct current
2	Quadrature current
3	Effective current
4	Effective voltage
5	Quadrature voltage
6	Phase current U
7	Phase current V
8	Phase current W

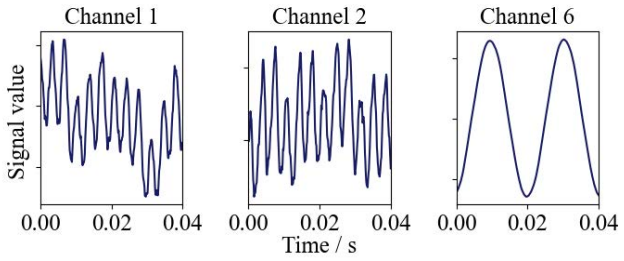


Figure 6. Overview of different derivatives of the current signal. Channel 1: direct current, Channel 2: quadrature current and Channel 6: phase current. All signals are in an internal normalization and thus unit free.

In contrast to vibration and acoustic signals, the original three phase currents are alternating, which may negatively influence some fault detection approaches. Further, the signals of at least two phases would be needed to cover all necessary information. To counter this, also the current in D-Q-coordinates as well as the effective current, calculated by the inverter, are stored, which is briefly shown in Fig. 6. Note that the direct current is related to the magnetic field, while the quadrature current is related to the motors torque. In addition, also the effective and quadrature target voltage are stored, however which are highly quantized and therefore may be of limited relevance.

4. Application

In this section, the presented methodology outlined in section 2 is applied on the gearbox data.

Firstly, only the data recorded at nominal speed in the loading direction with idle torque is considered. Additionally, the running-in process is discarded from the data. A running-in process is particularly well known for gears and causes volatile system behavior in the data shortly after machine commissioning. This can be caused by deforming or breakage of the highest asperities on the tooth surfaces (Feng et al.,

Table 5. Numbered overview of considered feature ranking methods.

No.	Name	Modified
1	Cori-Score	
2	Cori-Score	×
3	MTR_C	
4	MTR_C	×
5	MTR_Z	
6	Spearman	

Table 6. Minimum, maximum and mean value of the average RMSE for prediction on the training dataset within cross-validation across all feature selection variations.

Algorithm	Minimum	Maximum	Mean
GP	1.2e-9	9.8e-9	4.6e-9
LR	0.0956	0.2178	0.1458
MLP	0.0345	0.1220	0.0529
RF	0.0114	0.0314	0.0167
SVM	0.0571	0.0991	0.0662

2019). A running-in process is estimated at 50 hours. Therefore, the initial 50 feature data points, approximately two days of measurements, are rejected. Subsequently, 20,296 features are computed from the 8 channels of each gearbox data. The feature data set is then divided into a training and a test dataset, using the data from the H45 gearboxes for training and the B045 for testing. The target is to select meaningful features using the data from the H45 gearboxes to train ML algorithms, as discussed in section 2.4, and to apply them on the data from the B45 gearbox to finally predict its HI. In the following, the application of the proposed feature processing and selection methodology presented in sections 2.2 and 2.3 is described.

Within feature processing, the feature data undergoes smoothing, where a window size of 15 points seems appropriate. This window size is also used to determine the initial feature data points $f_{i,j,init}$ to scale the data. For threshold preselection, α is set to 1 % and β to 85 %. A small value for α can be selected as the running-in process has been removed. The value for β is chosen to consider the strong and varying increase of feature values towards the end of life. Threshold preselection leads to the exclusion of 19,572 features.

To rank the features, the feature ranking methods discussed in section 2.3 are employed. Due to the positive experience with the Spearman correlation specifically regarding capturing the HI of a system in (Aimiyekagbon, Bender, & Sextro, 2021), an additional version is employed, where the Spearman correlation is used for evaluating the monotonicity. An overview of the feature ranking methods is given by Tab. 5, where the additional versions are marked as modified. In the following, the numbers assigned in Tab. 5 are used as representatives for the mentioned ranking methods.

Lastly, for the correlation analysis to reject highly correlated features, the threshold value for the correlation coefficient is set to 0.98. A high value is chosen to remove strongly correlated features, thereby leave room for selection based on feature ranking. The selected threshold leads to the exclusion of 273 of 724 features. Subsequently, the top 5 ranked features are selected from the remaining 451 features, standardized and utilized for training and testing. A shuffle split with 5 splits is employed for cross-validation, as only two systems are given for training. To ensure the reproducibility of the results, the random seed is fixed. For optimization, 200 iterations are set.

For comparison purposes, additionally, the proposed feature selection methodology is replaced by feature ranking only. Feature selection consisting of feature ranking only represents the classical feature selection process, which is predominantly followed in the literature such as in (Carino et al., 2015; Nie et al., 2022). That means that out of the total of 20,296 features the top 5 ranked ones are used for training allowing a direct comparison with the proposed feature selection methodology.

5. RESULTS

The selected features, results and insights gained from further analysis are discussed in more detail in the following.

When inspecting the selected features, the channels 1, 2 and 3 show a significant higher relevance as they are selected 19, 24 and 10 times of 60 in sum respectively through feature selection. This observation leads to the conclusion that the current in D-Q-coordinates is particularly suitable for predicting the system's condition in contrast to the phase current. As assumed, the effective and quadrature voltage is of minor importance. Further, it can be observed that abrupt changes, reversed direction of feature progression and large differences in the endpoints between same features of the train and the test set cause confusion in prediction.

The minimal, maximal and mean RMSE of the predictions on the training data within cross validation is shown in Tab. 6 and Fig. 7 presents the prediction errors from predicting the HI of the gearbox B045. Primarily, all algorithms show low error values on the training dataset, which in combination with the results in Fig. 7 indicates, that some algorithms generalize better (RF) than other (MLP). MLP and SVM generate the highest RMSE, probably increased by the small amount of training data. SVM performs better evaluating the selected features from feature ranking only, although the predictions get particularly worse towards the end of life. The full potential of the MLP may not be exploited, as the iterations during its training and optimization are both limited to 200. In addition, the layer size and depth is not varied during optimization.

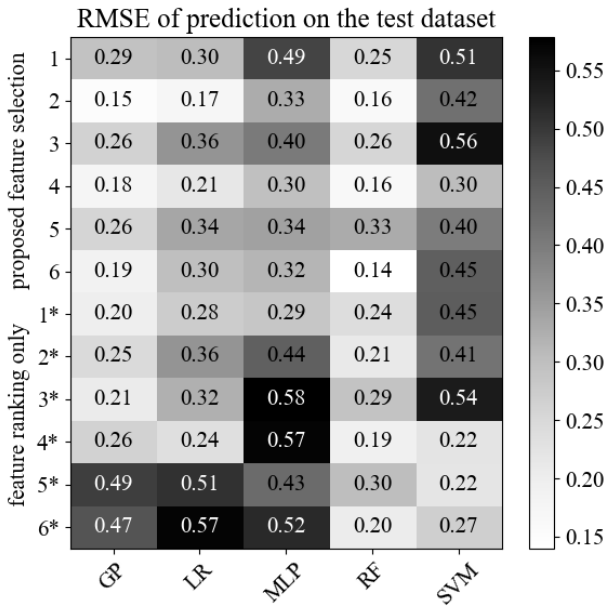


Figure 7. Results based on the estimation of the HI of gearbox B45, which provides the test data. The results are marked with an asterisk "*" when classical feature selection is applied and without when the proposed feature selection methodology is utilized.

The best performing algorithms are GP and above all RF. Especially noticeable is that the proposed selection method performs most effectively in combination with Spearman correlation for ranking the features as can be seen in rows 2, 4, and 6. To summarize, the best RMSE is reduced from 0.19 to 0.14 by around 26 % when the proposed feature selection method is utilized. The worst RMSE is reduced from 0.58 to 0.56 by around 3 %, whereby results with an RMSE of 0.5 and higher occur 6 versus 2 times and an RMSE of 0.4 and higher occurs 12 versus 7 times. Results with an RMSE of 0.2 and lower appear 2 versus 7 times. This indicates a higher robustness capabality using the proposed selection method.

The RF achieves the best result with an RMSE of 0.14 in row 6, where feature ranking within the proposed feature selection methodology is applied by assessing the Spearman correlation. The hyperparameters of the RF set by the hyperparameter optimization are given in Tab. 7 including a brief description. The selected features are shown in Fig. 8 where the feature values are plotted over the HI. The features are briefly described in Tab. 8. For detailed information on feature calculations, reference is made to the official documentation of tsfresh (Christ, Maximilian and Braun, Nils and Neuffer, Julius, 2016).

The prediction of the HI for the B45 gearbox data, generated by the RF trained on the presented features, is visualized in Fig. 9. The horizontal axis represents the actual HI values, while the vertical axis represents the predicted HI values. The diagonal line running from (1,1) to (0,0) represents the ideal prediction. The prediction of the test system shows a certain variance of the points, especially in the ranges 0.9 to 0.5 and 0.3 to 0.1 of the actual HI. The underestimated HI in the range 0.9 to 0.5 can be explained by the stronger gradient observed for the features 1 and 5. The overestimated HI in the range 0.3 to 0.1 is possibly caused by feature 2.

Table 7. Hyperparameter values and descriptions for the RF model set by the optimization algorithm.

Hyperparameter	Value	Description
n_estimators	149	Number of decision trees in the ensemble.
max_features	sqrt	Maximum number of features used to determine the best split. Here it is the square root of the number of features.
max_depth	27	Maximum depth of a single decision tree.
min_samples_split	1e-6	Minimum number of observations required to split a node in the decision trees. This number is defined by a fraction of the total number of observations.
min_samples_leaf	1e-6	Minimum number of observations required to form a leaf node. This number is defined by a fraction of the total number of observations.

Despite the observed variability, the prediction is deemed satisfactory considering the limited availability of training data and the structural differences between the systems for training and testing. The results presented underscore the ability of the proposed feature selection methodology in capturing the differences between the systems, especially in combination with the RF. Although certain challenges persist and continue to impact the overall results, the better results tend to align with the utilization of the proposed feature selection methodology, particularly shown in the upper half of the color map in Fig. 7.

6. Conclusion and Future Work

The effective use of available data is crucial, especially in scenarios characterized by data scarcity. The optimal use of available information is essential to improve the accuracy and reliability of prognostics and ensure efficient decision-making and resource allocation in the industry.

To tackle this challenge, comprehensive feature engineering with focus on feature selection is adopted, wherein the features are adapted to their initial values by scaling and feature selection is performed involving several successive steps. These steps encompass threshold-based preselection, feature ranking and cross-correlation analysis. Subsequently, training of ML-based models is conducted to predict the HI of the

Table 8. Description of the selected features obtained through the proposed feature selection methodology, wherein Spearman correlation was utilized for feature ranking.

Feature	Description
1	Value of the evaluated partial autocorrelation function at lag 6 of the quadrature current signal.
2	The highest order coefficient of a polynomial function the order 3 derived from the deterministic dynamics of the Langevin model, where 30 quantiles are used for averaging, based on the direct current data.
3	Complexity calculated by the Lempel-Ziv compression algorithm in the direct current data divided into 100 bins.
4	Custom feature explained in section 2.1, where direct current data is used. Bin 78 represents a frequency range from 26.12 to 26.86 Hz, where the FFT coefficients were aggregated by the mean.
5	The feature quantifies the maximum standard error of the linear trend over sections of length 5 in the direct current data.

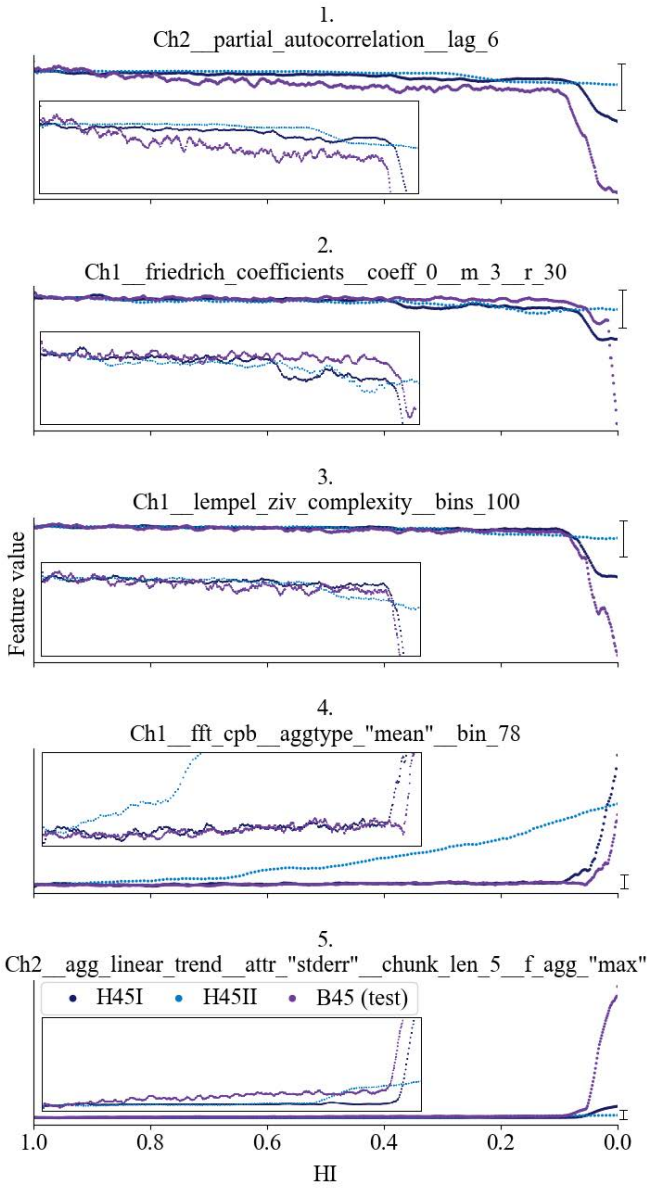


Figure 8. Best 5 features assessed by the proposed feature selection methodology based on the feature data from the H45 gearboxes, wherein the Spearman correlation was utilized for feature ranking. The boxes added indicate zoomed-in views of the features. The range is marked on the right edge. The feature values are unit free as they have been scaled.

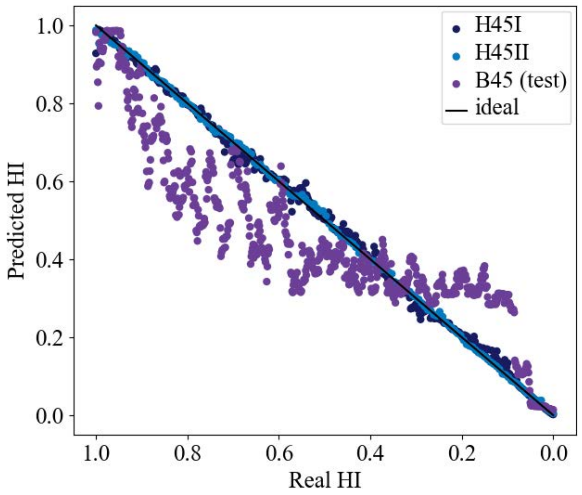


Figure 9. Test results generated by estimating the HI of the gearboxes H45I, H45II and B45, with the H45 gearboxes providing the training data and B45 gearbox the test data.

systems.

In order to evaluate the proposed methodology, a new dataset is introduced and utilized, which contains current, voltage and phase current data from run-to-failure experiments of gearboxes. The dataset is notable for considering two structurally different gearboxes and for addressing the challenge of data scarcity, as it is sourced from only three systems. The aim is to use the data from the two similar gearboxes to estimate and select features to infer the HI of the dissimilar gearbox over its entire operating time based on a ML algorithm. By publishing the novel dataset, other researchers are inspired to contribute to this specific problem setting.

It is observed that, the classical feature selection is able to select features capturing the degradation of the systems in some cases leading to an RMSE of 0.19 in the best case. However, the proposed feature selection methodology apparently supports overcoming system differences especially in combination with the RF by selecting appropriate features better leading to the best result overall with an RMSE of 0.14. Therefore, a great potential in applying the proposed methodology

to further problems in the field RUL-estimation is seen. It can enable more effective training based on ML training, as features are selected not only based on capturing the degradation of individual systems separately but also considering a common threshold for failure while avoiding redundancies.

The next research steps will include validation of the proposed methodology using further data or different test conditions to check the limitations, reliability and robustness of the results. The exploration of alternative methods, such as mutual information, should also be considered at the last step of the proposed feature selection process to replace the Pearson correlation analysis. These methods have the potential to enhance the methodology. Furthermore, the applicability of the proposed method to different types of gearboxes or even to other technical systems should be explored. This would contribute to demonstrating the scope and versatility of the proposed approach.

ACKNOWLEDGMENT

This research and development project is funded by the Ministry of Economy, industry, climate action and energy of the State of North Rhine-Westphalia (MWIKE) in the context of the Leading-Edge Cluster ‚Intelligent Technical Systems OstWestfalenLippe (it's OWL)' and supervised by Project Management Jülich (PtJ). The responsibility for the content of this publication lies with the author.

REFERENCES

Aimiyekagbon, O. K., Bender, A., & Sextro, W. (2021). On the applicability of time series features as health indicators for technical systems operating under varying conditions. *17. International Conference on Condition Monitoring and Asset Management (CM 2021)*.

Bender, A., & Sextro, W. (2021). Hybrid prediction method for remaining useful lifetime estimation considering uncertainties. *PHM Society European Conference*, *6*(1), 11. doi: 10.36001/phme.2021.v6i1.2843

Carino, J. A., Zurita, D., Delgado, M., Ortega, J. A., & Romero-Troncoso, R. J. (2015). Remaining useful life estimation of ball bearings by means of monotonic score calibration. In (pp. 1752–1758). doi: 10.1109/ICIT.2015.7125351

Chen, C., Xu, T., Wang, G., & Li, B. (2019). Railway turnout system rul prediction based on feature fusion and genetic programming. In (Vol. 151, p. 107162). doi: 10.1016/j.measurement.2019.107162

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, *307*, 72–77. doi: 10.1016/j.neucom.2018.03.067

Christ, Maximilian and Braun, Nils and Neuffer, Julius. (2016). *tsfresh.* `https://tsfresh.readthedocs.io`. (Accessed: March 07, 2024)

Feng, P., Borghesani, P., Chang, H., Smith, W. A., Randall, R. B., & Peng, Z. (2019). Monitoring gear surface degradation using cyclostationarity of acoustic emission. *Mechanical Systems and Signal Processing*, *131*, 199–221. doi: 10.1016/j.ymssp.2019.05.055

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press.

Goyal, D., Mongia, C., & Sehgal, S. (2021). Applications of digital signal processing in monitoring machining processes and rotary components: A review. In (Vol. 21, pp. 8780–8804). doi: 10.1109/JSEN.2021.3050718

Gram-Hansen, K. (1991). A bandwidth concept for cpb time-frequency analysis. In *[proceedings] icassp 91: 1991 international conference on acoustics, speech, and signal processing* (p. 2033-2036 vol.3). doi: 10.1109/ICASSP.1991.150803

Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, *240*, 98–109. doi: 10.1016/j.neucom.2017.02.045

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., & Shcherbatyi, I. (2021, October). *scikit-optimize: Sequential model-based optimization in python.* `https://zenodo.org/records/5565057`. Zenodo. (Last accessed: May 5, 2024)

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, *41*(14), 6371–6385. doi: 10.1016/j.eswa.2014.04.019

Li, Y., Huang, X., Gao, T., Zhao, C., & Li, S. (2023). A wiener-based remaining useful life prediction method with multiple degradation patterns. *Advanced Engineering Informatics*, *57*, 102066. doi: 10.1016/j.aei.2023.102066

Ly, C., Tom, K., Byington, C. S., Patrick, R., & Vachtsevanos, G. J. (2009). Fault diagnosis and failure prognosis for engineering systems: A global perspective. In (pp. 108–115). doi: 10.1109/COASE.2009.5234094

Nie, L., Zhang, L., Xu, S., Cai, W., & Yang, H. (2022). Remaining useful life prediction for rolling bearings based on similarity feature fusion and convolutional neural network. In (Vol. 44). doi: 10.1007/s40430-022-03638-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Thoppil, N. M., Vasu, V., & Rao, C. S. P. (2021). Health indicator construction and remaining useful life estimation for mechanical systems using vibration signal prognostics. *International Journal of System Assurance En-*

Table 9. Hyperparameter ranges for optimizing GP

Hyperparameter	Range
kernel	RBF, DotProduct, Matern, RationalQuadratic, WhiteKernel

Table 10. Hyperparameter ranges for optimizing MLP

Hyperparameter	Range	Distribution
activation	identity, logistic, tanh, relu	
solver	lbfgs, adam	
alpha	[1e-6, 1]	uniform
learning_rate	constant, adaptive, invscaling	
learning_rate_init	[1e-6, 1]	uniform

gineering and Management, *12*(5), 1001–1010. doi: 10.1007/s13198-021-01190-z

Yang, F., Habibullah, M. S., Zhang, T., Xu, Z., Lim, P., & Nadarajan, S. (2016). Health index-based prognostics for remaining useful life predictions in electrical machines. *IEEE Transactions on Industrial Electronics*, *63*(4), 2633–2644. doi: 10.1109/TIE.2016.2515054

Zhang, B., Zhang, L., & Xu, J. (2016). Degradation feature selection for remaining useful life prediction of rolling element bearings. In (Vol. 32, pp. 547–554). doi: 10.1002/qre.1771

BIOGRAPHIES

Alexander Loewen obtained his BSc and MSc in mechanical engineering at the Paderborn University, Germany. Since 2022, he is part of the Chair for Dynamics and Mechatronics at the Paderborn University. His research focuses on the automated training of ML based models for technical systems, particularly in anomaly detection, classification, and regression tasks under stationary operating conditions.

Peter Wissbrock is currently working as a researcher at the Innovation Department of Lenze SE. He obtained his BSc and MSc in electrical engineering at the OWL University of Applied Sciences and Arts, Germany. He is pursuing Ph.D. from Leibniz University Hannover, Germany in field of industrial analytics. He works in domain of drive train fault diagnosis, machinery anomaly detection and data acquisition infrastructure.

Amelie Bender studied mechanical engineering at RWTH Aachen University, Germany, and one semester abroad at the University of Newcastle, Australia. Since 2015 she is with the research group Dynamics and Mechatronics at Paderborn University, Germany. During her doctoral studies in mechanical engineering, her research focusses on condition monitoring of rubber-metal-bearings. She was awarded the academic degree Dr.-Ing. in 2021. As a team leader at the research group Dynamics and Mechatronics at Paderborn University, her research covers the topics condition monitoring, data analytics and reliability engineering.

Walter Sextro studied mechanical engineering at the Leibniz University of Hanover and at the Imperial College in London. Afterwards, he was development engineer at Baker Hughes Inteq in Celle, Germany and Houston, Texas. Back as research assistant at the University of Hanover he was awarded the academic degree Dr.-Ing. in 1997. Afterwards, he habilitated in the domain of mechanics under the topic Dynamical contact problems with friction: Models, Methods, Experiments and Applications. From 2004-2009 he was professor for mechanical engineering at the Technical University of Graz, Austria. Since March 2009 he is professor for mechanical engineering and head of the research group Dynamics and Mechatronics at the University of Paderborn.

APPENDIX

In Tabs. 9 to 12 the hyperparameter ranges are listed which where utilized for hyperparameter optimization of the regarding algorithm. For detailed information on hyperparameters, reference is made to the official documentation of scikit-learn (Pedregosa et al., 2011).

Table 11. Hyperparameter ranges for optimizing RF

Hyperparameter	Range	Distribution
n_estimators	[1, 200]	uniform
max_features	None, sqrt, log2	
max_depth	[1, 32]	uniform
min_samples_split	[1e-6, 1]	uniform
min_samples_leaf	[1e-6, 1]	uniform

Table 12. Hyperparameter ranges for optimizing SVM

Hyperparameter	Range	Distribution
C	[1e-2, 1e+3]	log-uniform
gamma	[1e-4, 1e+1]	log-uniform
kernel	linear, rbf	

Integrated design of negative stiffness honeycomb structures considering performance and operational degradation

Hyung-Do Kim[1], Taemin Noh[2], Young-Jin Kang[3], Nam-Ho Kim[4], and Yoojeong Noh[5*]

[1,2,5] *School of Mechanical Engineering Pusan National University, Busan, 46241, South Korea*

kimhd97@pusan.ac.kr
nohtm71@naver.com
yoonoh@pusan.ac.kr (Corresponding author)

[4] *Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, 32611, USA*

nkim@ufl.edu

[3] *Research Institute of Mechanical Technology, Pusan National University, Busan, 46241, South Korea*

zmanx@pusan.ac.kr

ABSTRACT

This study introduces an integrated framework for conceptualizing the design of negative stiffness honeycomb (NSH) structures, specifically considering the durability and performance of their unit cells. Unlike conventional energy-absorbing structures that rely on plastic deformation, NSH offers a promising alternative for reusable energy absorption (EA) and high initial stiffness, making it suitable for a wide range of engineering applications. The research considers the variability in characteristics of NSH based on the shape of the configured negative stiffness beam (NSB), selecting a single curved-beam unit cell as the focal point. Extensive testing, including quasi-static and cyclic compression tests, is conducted on NSH unit cell fabricated using polylactic acid/polyhydroxy alkenoate (PLA/PHA) filament, to analyze performance under stress and to assess degradation over time. Central to the study is the use of multi-objective optimization (MOO) to explore the trade-off between performance and operational durability, thereby emphasizing the significance of degradation in the design process. The results demonstrate the potential for NSH structures, particularly in terms of their reusability and efficiency, highlighting the viability of incorporating durability considerations in the early stages of design, especially for structures intended for additive manufacturing processes.

1. INTRODUCTION

NSH structures exhibit unique characteristics when compared to traditional hexagonal honeycombs. While hexagonal honeycombs effectively absorb energy through plastic deformation, they fall short in terms of reusability post-deformation. (Correa et al., 2015) NSH structures, composed of NSBs, stand out for their recoverable energy absorption, as highlighted by (Klatt et al., 2013; Correa et al., 2015), their high initial stiffness (Correa et al., 2015), and their capabilities in impact isolation (Shan et al., 2015; Debeau et al., 2018), creating opportunities for their use in many engineering fields.

Many studies have been conducted on the characteristics of such NSBs. Qiu et al. (2004) studied a bistable mechanism with a curved beam, whereas Klatt et al. (2013) demonstrated negative stiffness behavior and recoverable energy absorption through vertical axial compression in an additively manufactured structure with a curved beam. Correa et al. (2015) optimized the dimensions of NSH, achieving a structure with similar relative density and force threshold as traditional hexagonal honeycomb, but with better energy absorption per unit mass, closely matching the performance of the hexagonal honeycomb. Chen et al. (2021) showed that NSH, comprising curved beams of varying thicknesses, not only improved energy absorption per mass but also enhanced shock absorption and vibration isolation compared to uniform-thickness NSH. Zhang et al. (2021) proposed a lattice and hollow structure for the curved beam, showing better energy dissipation than conventional curved beams of the same volume. Liu et al. (2020) used machine learning methods to achieve enhanced results in curved beam thickness optimization. In addition, research on cylindrical structure (Wang et al., 2020), cubic structure (Ha et al., 2019), and composite negative stiffness structure (Chen et al., 2020) shows various negative stiffness structures and different features depending on the shape and dimensions of NSBs.

A key feature of negative stiffness structures like NSH, distinguishing them from other structures, is their reusability. The studies in Correa et al. (2015), Tan et al. (2019), and

Chen et al. (2020) show the properties of negative stiffness structures, such as force thresholds and energy absorption or dissipation under repeated compression. Chen et al. (2020) shows that the degree of reduction in force threshold for cyclic compression depends on the dimensions of the NSB's thickness. However, there has been limited research on quantifying the performance reduction of negative stiffness structures relative to the NSB dimensions under cyclic compression, which is crucial for predicting the operational end of life (EOL) of these structures.

To address this gap, we propose an integrated design framework that considers both the performance and operational aspects of negative stiffness structures like NSH, including performance degradation. In this study, we targeted the unit cell of NSH for design and manufactured it using PLA/PHA filament through 3D printing. To consider both performance and operational aspects, we conducted quasi-static compression tests and cyclic compression tests to acquire data. Based on this data, we developed a model to estimate the performance and EOL of the NSH unit cell according to its dimensions. Finally, through the Multi-objective Optimization (MOO) design process considering the estimated performance and EOL of the NSH unit cell, we not only confirmed the relationship between the structural performance and operational aspects but also provided insights into the design considering both aspects.

2. DESIGN OF EXPERIMENT

The unit cell of NSH, as illustrated in Figure 1, was employed in this study. The structure of the curved-beam is assumed to be based on Eq. (1). (Qiu et al., 2004) The design variables defined for this structure are the thickness (t) and central height (h) of the curved beam.

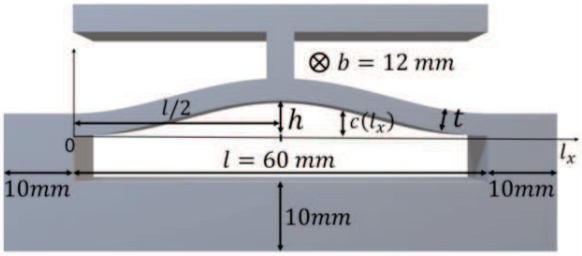


Figure 1. Geometry and dimension of unit cell of NSH

$$c(l_x) = \frac{h}{2}\left[1 - cos\left(2\pi\frac{l_x}{l}\right)\right] \qquad (1)$$

Considering that Zhakatayev et al. (2020) and Tan et al. (2019) have confirmed that the influence of the thickness (t) and height (h) of NSB on the strength, absorbed energy per unit mass, and force threshold of negative stiffness structure, and Qiu et al. (2004) have established a relationship between force-displacement of the curved-beam and geometric parameters as per Eq. (2),

$$F = \frac{3\pi^4 Q^2}{2} d_n \left(d_n - \frac{3}{2} + \sqrt{\frac{1}{4} - \frac{4}{3Q^2}}\right)\left(d_n - \frac{3}{2} - \sqrt{\frac{1}{4} - \frac{4}{3Q^2}}\right) \qquad (2)$$

Klatt et al. (2013) observed that negative stiffness initiates when the numerical value of $Q(= h/t)$ reaches 1.5, when Q exceeds 2.31, the bi-stable characteristics become evident. Therefore, t and h can be considered as important design factors for the negative stiffness structure like Fig 1.

Therefore, we defined the range of t as $1.2[mm] \leq t \leq 3.2[mm]$ and h as $1.2[mm] \leq h \leq 6.4[mm]$. Subsequently, we sampled samples using the design of experiment (DOE) method to train and test the surrogate models and classification models for the characteristics of NSH unit cell, which will be discussed later in section 4 and 5. First, 25 samples were sampled for the training data using the full factorial design (FFD) method. For the test data, 10 samples were sampled through the optimal Latin hypercube design (OLHD) method. The results are illustrated in Figure 2.

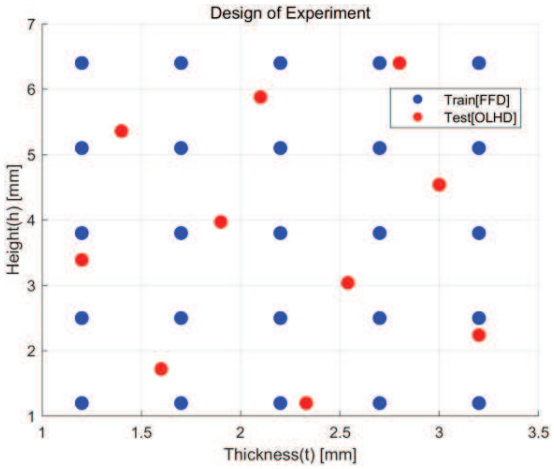


Figure 2. Result of design of experiment

3. FABRICATION AND TEST

We manufactured NSH unit cells with dimensions obtained through the DOE process using a fused filament fabrication (FFF) 3D printer and PLA/PHA filament. Liu et al. (2023) demonstrated that variations in manufacturing features, such as building direction, fill pattern, and wall layers influence printing quality and performance of the NSH cell through the FFF method. Therefore, we considered three different infill angles for 3D printing. We utilized Simplify 3D software for 3D printing, and detailed printing settings can be found in Table 1.

3.1. Material Properties

In this study, Colorfabb's PLA/PHA filament was utilized for fabricating NSH unit cell. Research conducted by Morettni et al. (2022), Letcher & Waytashek et al. (2014), Zouaoui et al. (2021), and Gonabadi et al. (2020) have confirmed that the physical properties of FFF 3D printing can

vary depending on manufacturing parameters such as infill angle or pattern. Therefore, to account for these manufacturing characteristics, five specimens were printed with three different infill angles (0°, 45°, and 90°) to assess the physical properties of the PLA/PHA filament through ASTM D638. An example of specimens is depicted in Figure 3, and the result of ASTM D638 are presented in Table 2.

Table 1. 3D Printing setting.

Nozzle Temperature	210 ℃
Bed Temperature	60 ℃
Infill Density	100 %
Infill Pattern	Rectilinear
Infill Angle	[0°, 45°, 90°]
Layer height	0.2 mm
Printing Speed	50 mm/s
Cooling Fan Speed	100 %
Building direction	Flat
Material	PLA/PHA

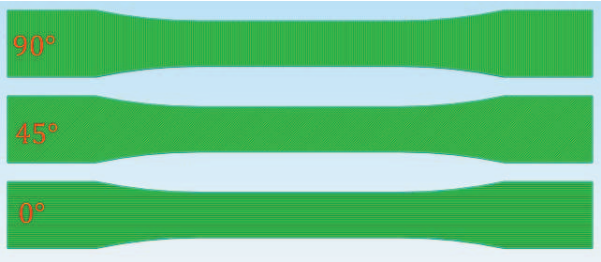


Figure 3. ASTM D638 specimen with three angles of infill

In Table 2, the average values and standard deviations of the ASTM D638 test results show that the average values of Young's modulus, yield strength, and elongation decreases as the infill angle increases from 0° to 45° and 90°. This is because as the infill angle increases, the force applied to the specimen and the direction of the stacked filament become more closely perpendicular. Therefore, when manufacturing the unit cells of NSH through 3D printing, we set the infill angle to 0° and produced 5 unit cells of NSH per sample. An example is illustrated in Figure 4.

Table 2. Material properties according to infill angles

Infill angle	Young's Modulus [GPa]	Yield Strength [MPa]	Elongation [%]	Poisson's Ratio
0°	2.84 (0.059)	52.11 (0.548)	5.84 (1.629)	0.34 (0.005)
45°	2.59 (0.019)	37.86 (0.833)	5.75 (1.561)	-
90°	1.99 (0.062)	16.48 (1.250)	1.65 (0.366)	-

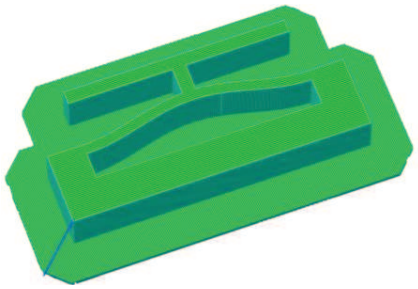


Figure 4. A unit cell for NSH with 0° infill angle

3.2. Compression Test for Data Acquisition

Quasi-static compression tests and cyclic compression tests were conducted to acquire experimental data, considering the performance and operational aspects of the NSH's unit cell. In both tests, compression was applied by inducing a displacement of $2h$ to the unit cell of NSH. The compression test equipment comprised a JSV-1000 stand and a HF-100 force gauge. Additionally, consistent compression test conditions were maintained throughout by securing both ends of the structure using a support structure, as depicted in Figure 5.

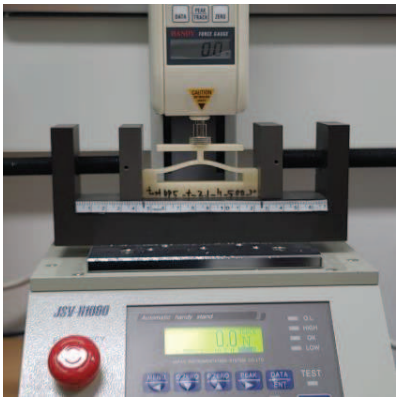


Figure 5. Compression test equipment and environment

However, different types of NSH unit cells were utilized in the two types of tests, as shown in Figure 6. The structure depicted in Figure 6 represents a configuration designed for quasi-static compression test. Unlike Figure 1, an additional structure is incorporated at the compression center of the T-shaped support to minimize asymmetric buckling mode in the curved-beam behavior. Conversely, for the cyclic compression test, these additional structures may interfere with the cyclic compression process, hence a configuration similar to Figure. 1 was employed.

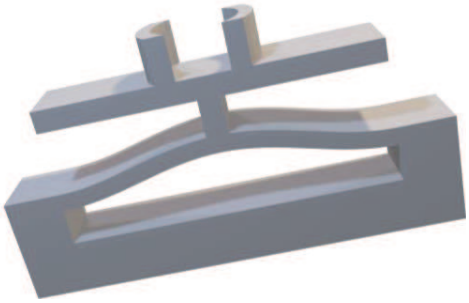


Figure 6. Shape of the NSH unit cell used in quasi-static compression test

3.2.1. Quasi-static Compression Test

To assess the performance aspect of NSH unit cell, a quasi-static compression test was conducted at a speed of 10 mm/min. The obtained force-displacement data were preprocessed using a moving average filter to generate five force-displacement curves for each sample, as illustrated in Figure 7.

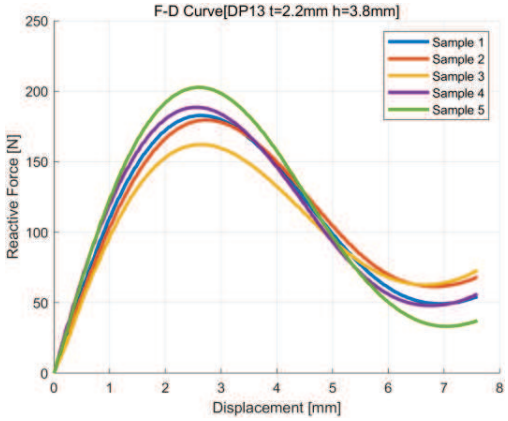


Figure 7. Force-displacement curve for quasi-static compression test

Then, specific energy absorption (SEA) was obtained by dividing Eq. (4) by Eq. (3), with the average SEA value designated as the representative value for the corresponding sample.

$$m = \rho b\left[\left(\int_0^l (c(l_x)+t)dw\right)-\left(\int_0^l c(l_x)\,dw\right)\right] \quad (3)$$

$$EA = \left(\int_0^{2h} f(w)dw\right) \quad (4)$$

These data were also used to examine the occurrence of negative stiffness for five samples of each design point employed in the experiments, as detailed in Section 5. The occurrence of negative stiffness was assessed using Eq. (5), as established by Qiu et al. (2004), and Eq. (6) based on the force-displacement data.

$$w_{mid} = \frac{4}{3}h \quad (5)$$

$$\begin{cases} max_{w<w_{mid}}(f(w)) - f(w_{mid})) > 0 \rightarrow Negative\ stiffness \\ otherwise \rightarrow Non-Negative\ stiffness \end{cases} \quad (6)$$

This allowed us to classify whether negative stiffness occurred based on w_{mid} in the force-displacement curve.

3.2.2. Cyclic Compression Test

In this experiment, 30 cycles of compression were repeatedly applied at a speed of 60 mm/min. The force-displacement data obtained underwent the same data preprocessing as the quasi-static compression test. The average force-displacement curve for each sample is depicted in Figure 8. Using mean force-displacement data, EA for each cycle was calculated using Eq. (4); mean force-displacement data was also utilized as the health index (HI) for estimating the end of life (EOL), a topic discussed in detail in Section 6.

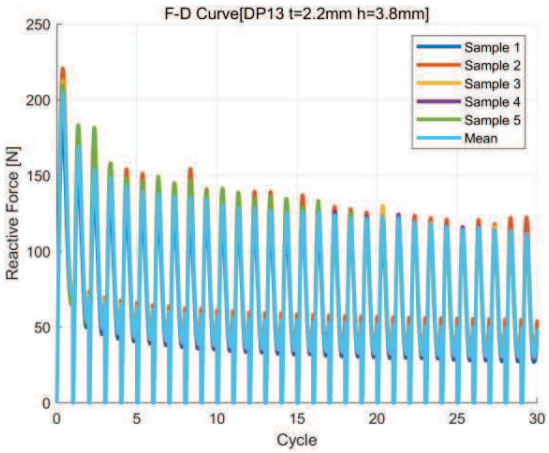


Figure 8. Reactive force for each cycle of the cyclic compression test

4. SURROGATE MODEL

A surrogate model replaces a high-cost test-based or simulation model with a relatively low-cost alternative. By creating a surrogate model for a specific factor of interest, predictions can be made without the need for costly tests or simulations for any given sample. In this study, the Kriging method, implemented in the PIAnO 2024 software, was used to develop a surrogate model for the performance and operational factors of the NSH unit cell. This approach enabled the prediction of the values of these factors for a specific design point.

4.1. Kriging

Kriging is one of the most widely used methods for constructing a surrogate model or metamodel, also known as Gaussian process regression. Based on the references to

Forrester et al. (2008) and Kim et al. (2017), the explanation of Kriging would be as follows: In Kriging, the predicted output of a Kriging model is typically represented as Eq. (7).

$$\hat{y}(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta} + \delta(\mathbf{x}) \tag{7}$$

Here, $\mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}$ represents the global function, and $\delta(\mathbf{x})$ represents the local departure. In our study, $\mathbf{x}$ denotes the dimensions of the NSH unit cell, such as t and h, while $\hat{y}$ represents the value we want to predict, such as SEA. We have defined the correlation function for two points ($\mathbf{x}, \mathbf{x}'$) as shown in Eq. (8).

$$\Gamma(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{ndv} \exp\left(-p_i |x_i - x_i'|^{1.5}\right) \tag{8}$$

where ndv indicates the number of design variables, and p_i represents the parameter of the correlation function. Other types of covariance functions can be found in Rasmussen & Williams (2006) and Xu (2020). Therefore, the correlation matrix is expressed as shown in Eq. (8), and the correlation between the point $\mathbf{x}$ to be predicted and the observed points is expressed as shown in Eq. (9).

$$\boldsymbol{\Gamma} = \begin{pmatrix} \Gamma(\mathbf{x_1}, \mathbf{x_1}) & \cdots & \Gamma(\mathbf{x_1}, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \Gamma(\mathbf{x}_N, \mathbf{x_1}) & \cdots & \Gamma(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \tag{9}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \Gamma(\mathbf{x_1}, \mathbf{x}) \\ \vdots \\ \Gamma(\mathbf{x}_N, \mathbf{x}) \end{pmatrix} \tag{10}$$

To estimate the parameters of the Kriging model, $\boldsymbol{\theta}$, s^2, and $\mathbf{h}$, we use maximum likelihood estimation (MLE). So, the logarithmic likelihood can be expressed as Eq. (11).

$$\begin{aligned} &\ln\left(L(\mathbf{y}|\boldsymbol{\theta}, s^2)\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(s^2) - \frac{1}{2}\ln(|\boldsymbol{\Gamma}|) \\ &- \frac{(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})}{2s^2} \end{aligned} \tag{11}$$

Taking the derivatives of Eq. (11) for $\boldsymbol{\theta}$ and s^2 respectively, and setting them to zero, yields the estimation results via MLE as shown in Eqs. (12) and (13).

$$\hat{\boldsymbol{\theta}} = (\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{G})^{-1}(\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{y}) \tag{12}$$

$$\hat{s}^2 = \frac{(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})}{2N} \tag{13}$$

The parameter h is determined by substituting Eqs. (12) and (13) into Eq. (11), and the resulting value is maximized by the optimization algorithm (Differential evolution, DE), as expressed in Eq. (14).

$$p = argmax\left[-\frac{N}{2}\ln(\hat{s}^2) - \frac{1}{2}|\boldsymbol{\Gamma}|\right] \tag{14}$$

Given a vector $\hat{\mathbf{y}} = [\mathbf{y}^{\mathrm{T}}, \hat{y}\,]^{\mathrm{T}}$, which includes the new predicted value $\hat{y}$ at $\mathbf{x}$, the correlation matrix can be written as Eq. (15).

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} \boldsymbol{\Gamma} & \gamma \\ \gamma^{\mathrm{T}} & 1 \end{pmatrix} \tag{15}$$

Based on this, we obtain the logarithmic likelihood, as shown in Eq. (16).

$$\begin{aligned} &\ln(L) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\hat{s}^2) - \frac{1}{2}\ln\left(|\hat{\boldsymbol{\Gamma}}|\right) \\ &- \frac{(\hat{\mathbf{y}} - \mathbf{G}\hat{\boldsymbol{\theta}})^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}^{-1}(\hat{\mathbf{y}} - \mathbf{G}\hat{\boldsymbol{\theta}})}{2\hat{s}^2} \end{aligned} \tag{16}$$

Differentiating Eq. (16) with respect to $\hat{y}$ and setting it to zero, the final output of a Kriging model is expressed as Eq. (17):

$$\hat{y}(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\hat{\boldsymbol{\theta}} + \boldsymbol{\gamma}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}(\mathbf{y} - \mathbf{G}\hat{\boldsymbol{\theta}}) \tag{17}$$

4.2. SEA Prediction Model

We formed a surrogate model for SEA to consider the performance aspect of the NSH unit cell. To do this, we first performed a quasi-static compression test on the 25 samples collected by the FFD method, with 5 samples per test point. The average SEA results for each sample were successfully obtained.

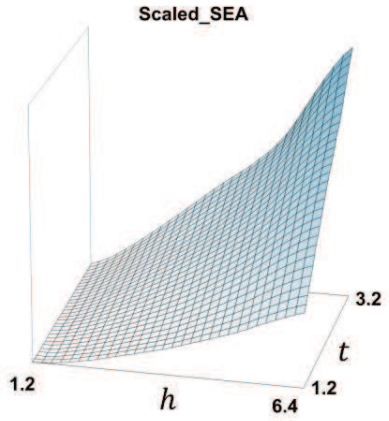


Figure 9. Response surfaced of Scaled SEA_{mean}

Prior to creating the surrogate model using 25 datasets, we set the design variables of the NSH unit cell, t and h, as the inputs for the surrogate model, with SEA_{mean} as the output. Both input and output data were scaled to have values between 0 and 1 using min-max scaling. Finally, we set the global function type to constant, and the results of this surrogate model are depicted in Figure 9. As shown in Figure 9, SEA_{mean} tends to increase as the values of the design variables t and h increase. The root mean square error (RMSE) for this surrogate model was computed using Eq. (18) with 10 test data points, resulting in an RMSE of 0.0276.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{pred,i} - y_{true,i})} \quad (18)$$

4.3. Energy Absorption over Cycles Prediction Model

Similar to Section 4.2, Kriging was utilized to generate surrogate models for predicting EA over the compression cycle using average force-displacement data from about 25 samples collected via FFD method.

Initially, EA for each compression cycle was computed using Eq. (19):

$$EA_{cycle} = \left(\int_0^{2h} f_{cycle}(w)dw\right) \quad (19)$$

Following this, it was assumed that there was no degradation in the NSH unit cell prior to cyclic compression test, and Eq. (20) was used to scale based on 1 cycle of EA as a reference.

$$EA_{sc} = \frac{EA_{cycle}}{EA_{\,1\,cycle}} \quad (20)$$

An example of EA_{sc} is shown in Figure 10, where it is crucial to note that for any sample, EA_{sc} is 1 at 1 cycle. EA_{sc} was used as the HI for estimating EOL.

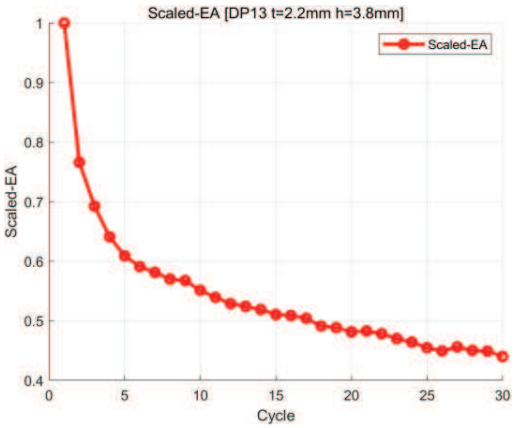


Figure 10. Scaled EA over cycles

The EA_{sc} was further processed by Eq. (21) for samples with a EA_{sc} exceeding 0.9 at 30 cycles:

$$\begin{cases} EA_{sc} = 1, \quad for \ \ cycle = 1 \\ EA_{sc} = EA_{sc-1} - |EA_{sc} - EA_{sc-1}|, otherwise \end{cases} \quad (21)$$

The input data, consisting of t and h, was used to train the model, aiming to predict EA_{sc} for a specific cycle. Unlike the surrogate model for SEA_{mean}, only min-max scaling was applied to the input data, and a simple quadratic function was utilized as the global function to construct the surrogate model. The corresponding response surface for this is shown in Figure 11.

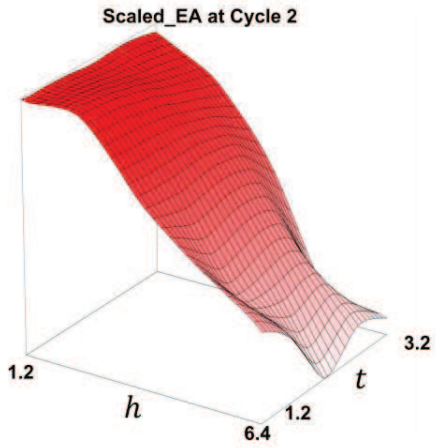


Figure 11. Response surfaced of EA_{sc} at cycle 2

The surrogate model predicts EA_{sc} for each cycle, which is then utilized to estimate the EOL for NSH unit cells. This will be discussed in detail in Section 6.

5. CLASSIFICATION MODEL

A classification model was developed based on the findings discussed in Section 3.2.1, where the design of NSH unit cells exhibits negative stiffness depending on certain design variables. Previous studies by Shahan et al. (2012), Morris et al. (2018), and Matthews et al. (2016) demonstrated that a set-based approach using the Bayesian network classifier method can be used to explore the boundaries of the design space and identify designs that meet specific performance criteria. Based on this, we utilized the Bayesian classifier as a classification model to determine the presence of negative stiffness. We formed the classification model using the results from quasi-static compression tests on 25 samples. Furthermore, this classification model was used as a constraint in the MOO design process, which will be discussed in detail in Section 7.

5.1. Bayes classifier

The results from all five test points in the quasi-static compression test were incorporated into the Bayesian classifier model. Specifically, the prior probability, as defined by Eq. (22) from Shahan et al. (2012), was established based on the frequency of occurrence of negative stiffness.

$$\begin{cases} P(c_{NS}) = \dfrac{N_{NS} + 1}{N + 2} \\ P(c_{NNS}) = \dfrac{N_{NNS} + 1}{N + 2} \end{cases} \quad (22)$$

For the likelihood, multivariate kernel density estimation was employed as described by Scott (2015) and can be expressed using Eq. (23):

$$\begin{cases} P(\mathbf{x}|c_{NS}) = \frac{1}{N_{NS}\beta_{1,NS}\cdots\beta_{N_d,NS}}\sum_{i=1}^{N_{NS}}\left\{\prod_{j=1}^{N_d}K\left(\frac{x_j - x_{ij}}{\beta_j}\right)\right\} \\ P(\mathbf{x}|c_{Non-NS}) = \frac{1}{N_{NNS}\beta_{1,NNS}\cdots\beta_{N_d,NNS}}\sum_{i=1}^{N_{NNS}}\left\{\prod_{j=1}^{N_d}K\left(\frac{x_j - x_{ij}}{\beta_j}\right)\right\} \end{cases} \quad (23)$$

The Gaussian kernel K is used, and the bandwidth β values are calculated using Eq. (24):

$$\begin{cases} \beta_{j,NS} = s_j\left\{\frac{4}{(N_d+2)N_{NS}}\right\}^{1/(N_d+4)} \\ \beta_{j,NNS} = s_j\left\{\frac{4}{(N_d+2)N_{NNS}}\right\}^{1/(N_d+4)} \end{cases} \quad (24)$$

The posterior probabilities for the two classes are given by Eq. (25).

$$\begin{cases} P(c_{NS}|\mathbf{x}) = P(c_{NS})P(\mathbf{x}|c_{NS}) \\ P(c_{NNS}|\mathbf{x}) = P(c_{NNS})P(\mathbf{x}|c_{NNS}) \end{cases} \quad (25)$$

$$\lambda_1 P(c_{NS})P(\mathbf{x}|c_{NS}) - \lambda_2 P(c_{NNS})P(\mathbf{x}|c_{NNS}) > 0 \quad (26)$$

Then, the decision rule for class for classifying a sample regarding the occurrence of negative stiffness is defined by Eq. (26) below. According to the study by Shahan et al. (2012), it has been confirmed that the loss factor λ_1 and λ_2 can shift the decision boundary of the classifier. Therefore, setting $\lambda_1 = 0.66$, $\lambda_2 = 0.34$ accounts for cases where negative and non-negative stiffness may occur simultaneously in the samples. This setting allows the classification of such samples into the class indicating the occurrence of negative stiffness. With $\lambda_1 = 0.66$ and $\lambda_2 = 0.34$, the difference between the two posterior probabilities is illustrated in Figure 12.

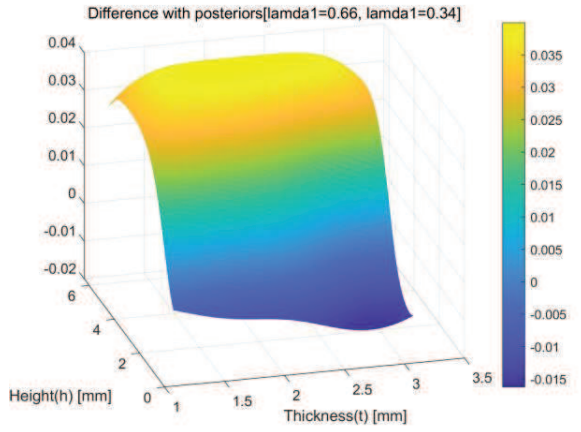


Figure 12. Difference between the two posterior probabilities

6. PREDICTING EOL OF NSH UNIT CELL

In order to consider the operational aspects of the NSH unit cell, prognostics methods were utilized to estimate the EOL. According to Kim et al. (2017), prognostics methods can be categorized into physics-based and data-driven approaches. Kim et al. (2017) also introduced nonlinear least square (NLS), Bayesian method (BM), and particle filter (PF) within physics-based prognostics. In this study, the NLS method demonstrated by Kim et al. (2017) was used to estimate the EOL by considering degradation, as shown in Figure 10, through EA_{sc}, which serves as the HI of the NSH unit cell.

First, to estimate the EOL via NLS, the degradation equation was defined as Eq. (27):

$$\varphi = exp\left(-|\xi_1| \cdot (cycle - 1)^{\xi_2}\right) \quad (27)$$

The parameters ξ_1 and ξ_2 were estimated using the 'lsqnonlin' function in MATLAB R2023b, employing the Levenberg-Marquardt method. To consider the uncertainty of the estimated model parameters in NLS, the 95% confidence intervals for the model parameters were obtained from 1.0E7 random sampling from the multivariate t-distribution using Eqs. (28) and (29), with degrees of freedom $N - N_p + 1$. Here, Eq, (28) represents the variance of noise in measured data, and Eq. (29) represents the variance of estimated model parameters.

$$s_n^2 = \frac{\{\mathbf{y} - \boldsymbol{\varphi}\}^{\mathrm{T}}\{\mathbf{y} - \boldsymbol{\varphi}\}}{N - N_p} \quad (28)$$

$$\mathbf{M}_{\xi} = s_n^2[\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}]^{-1} \quad (29)$$

The challenge in EOL estimation lies in determining amount of EA_{sc} data needed to estimate the EOL using the surrogate model from Section 4.3, and how to estimate parameters ξ_1 and ξ_2 using NLS. To address this, a model was developed to predict EA_{sc} for design variables t and h across 2 to 15 cycles using the Kriging model from Section 4.3. For the 10 test data, the EA_{sc} data estimated by the surrogate model from 3 to 15 cycles was progressively added, calculating the median of the confidence interval of NLS and the mean RMSE of the actual experimental data. The results of mean RMSE are depicted in Figure 13.

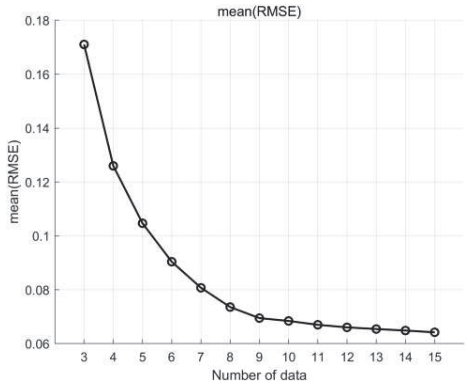


Figure 13. Mean RMSE by number of data

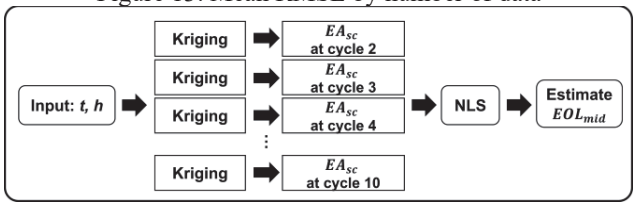


Figure 14. EOL Estimation Process

It was observed that utilizing more than 10 cycles of data predicted from surrogate model (i.e., beyond 1-10 cycles) did not significantly affect the estimation error in degradation estimation via NLS. Consequently, EA_{sc} values were estimated for 2-10 cycles through a surrogate model as shown in Figure 11, considering that at 1 cycle, the EA_{sc} value is consistently 1 across all samples. This process is illustrated in Figure 14. Therefore, this process was utilized to estimate the EOL of the NSH unit cell, and the median of the EOL confidence interval (EOL_{mid}) was used in the MOO design process, which will be discussed in detail in Session 7.

7. MULTI-OBJECTIVE OPTIMIZATION DESIGN

A MOO design was implemented to address both the performance and operational aspects of the NSH unit cell. SEA_{mean} was considered for the performance aspect, while the estimated EOL_{mid} served as the objective function for the operational aspect. Constraints included the strain of the curved beam, the threshold for HI, and the presence or absence of negative stiffness. The problem was formulated accordingly, and the results of the MOO design were analyzed using the NSGA-2 optimization algorithm (Deb et al., 2002), implemented in the PIAnO 2024 software. The overall flowchart is depicted in Figure 15.

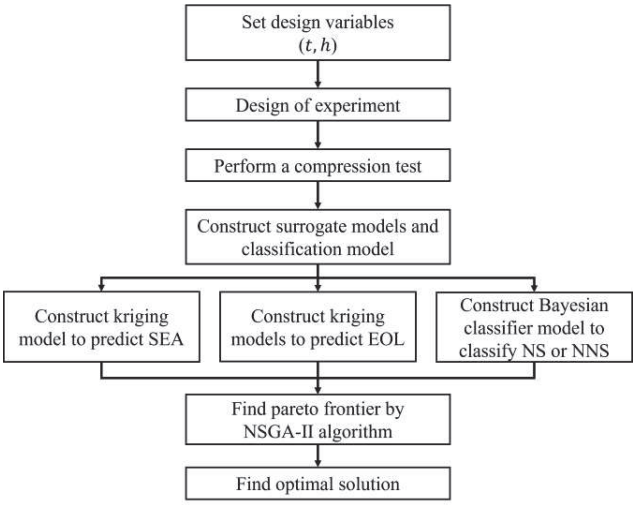


Figure 15. Flowchart of multi-objective optimization design

7.1. Problem Formulation

For the MOO design, the problem formulation is defined as Eq. (30). Initially, the surrogate model for SEA_{mean}, as discussed in Section 4.2, was employed to address the performance aspect of the NSH unit cell. Maximizing the SEA_{mean} implied enhancing the capacity of the unit's curved beam to absorb energy relative to its mass. Subsequently, the EOL_{mid} estimated through the approach outlined in Section 6, was considered for the operational aspect. To ensure comparability in scale between the scaled SEA_{mean} by min-max scaling and the estimated EOL_{mid}, we utilized the EA_{sc} obtained from cyclic compression tests from 1 to 30 cycles on the 25 samples extracted using the FFD method to estimate EOL_{mid}. Based on this estimation, we performed min-max scaling on the estimated EOL_{mid}. At this point, it was assumed that the EOL_{mid} from 25 samples provides sufficient information about EOL_{mid} for the entire design space.

The first constraint was defined using the maximum strain, determined from the mean elongation when the infill angle is 0°. The second constraint was defined as the occurrence of negative stiffness, where $\lambda_1 = 0.66$ and $\lambda_2 = 0.34$. The threshold for EA_{sc} as HI was assumed to be 0.7, indicating that the structure has degraded to 30% of its original performance.

$$\begin{aligned} &Find && t, h \\ &maximize && f(\mathbf{x}) \\ &&& = SEA_{sc,mean}(t_{sc}, h_{sc}) \\ &&& + EOL_{sc,mid}(t_{sc}, h_{sc}) \\ &subject\ to && 2\pi^2 \frac{th}{l^2} < 0.0584 \\ &&& 0.66P(c_{NS})P(t,h|c_{NS}) \\ &&& - 0.34P(c_{NNS})P(t,h|c_{NNS}) > 0 \\ &&& Threshold = 0.7 \\ &&& 1.2\ mm \le t \le 3.2\ mm \\ &&& 1.2\ mm \le h \le 6.4\ mm \end{aligned} \tag{30}$$

7.2. Result of Multi-Objective Optimization Design

In the MOO design process, we considered an initial design point for the NSH unit cell with $t = 2.2\ mm$ and $h = 3.8\ mm$. NSGA-2 was employed as the optimization algorithm in the PIAnO 2024 software, with settings summarized in Table 3.

Table 3. The settings for NSGA-2

Population Size	100
Crossover Rate	0.9
Mutation Rate	0.5
Maximum Number of Generations	250

The results are displayed in Figure 16, where the lower constraints pertain to the condition for the occurrence of negative stiffness, while the upper constraints relate to the maximum strain. When plotting the Pareto frontier for the objective function, it appears similar to Figure 17.

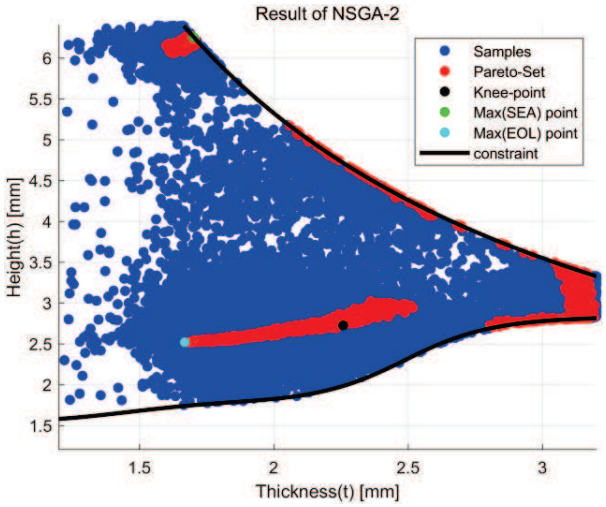


Figure 16. Optimum results using NSGA-2

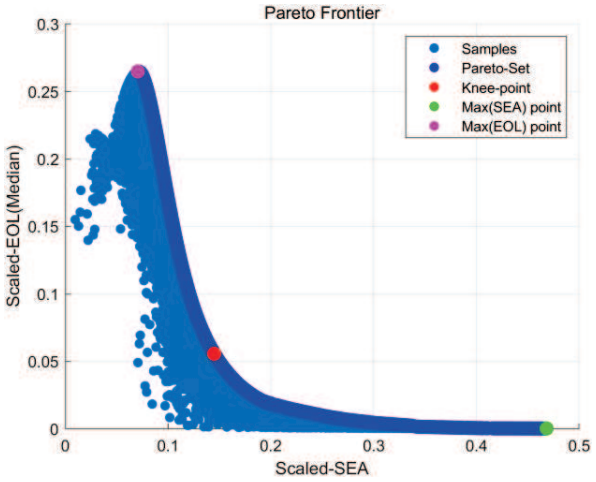


Figure 17. Results of Pareto frontier

Figure 17 illustrates that, despite aiming to maximize the objective functions, EOL and SEA exhibit an inverse relationship within the Pareto optimal set. As EOL increases, SEA decreases, and vice versa. Therefore, to identify the knee point, a horizontal line was extended from the point with the maximum SEA_{mean}, and a vertical line from the point with the maximum EOL_{mid}. The knee point was determined as the intersection of these lines, selected as the closest point from the Pareto optimal set. The dimensional information for the points with the maximum SEA_{mean}, maximum EOL_{mid}, and the knee point is provided in Table 4.

Table 4. Dimensions for three points

	Max (SEA_{mean})	Max (EOL_{mid})	Knee Point
t [mm]	1.702	1.669	2.258
h [mm]	6.255	2.52	2.725

After manufacturing, quasi-static compression tests and cyclic compression tests were conducted for these three points, as discussed in Section 3. All three points satisfied maximum strain constraint as defined in Eq. (29) and exhibited negative stiffness in five samples per point during the test. The results for SEA_{mean} and EOL_{mid} were summarized in Tables 5 and 6.

Table 5. SEA_{mean} results of 3 points

	Initial point	Max (SEA_{mean})	Max (EOL_{mid})	Knee point
True [mJ/g]	425.96	641.06	116.05	207.34
Predict [mJ/g]	-	760.93	128.27	246.42

Table 6. Estimated EOL_{mid} results of 3 points

	Initial point	Max (SEA_{mean})	Max (EOL_{mid})	Knee point
Estimated EOL_{mid} [95% C.I] by proposed method (Cycle)	-	1.49 [1.47 ,1.52]	102.08 [50.21, 266.95]	22.56 [14.78, 39.04]
Estimated EOL_{mid} from 30 cycles of true test data (Cycle)			127.91	31.44
True EOL (Cycle)	2.90	1.48	-	-

According to Table 5, the SEA_{mean} value at the point where it reaches its maximum is approximately 50.5% higher than the initial design point, as predicted by the surrogate model.

Table 6 displays the estimated EOL_{mid} results for the three points. The median result estimated by the proposed method was compared with actual compression test data collected over 30 cycles. However, it is important to note that since cyclic compression test data are available only up to 30 cycles, the EOL beyond this point cannot be accurately determined. Therefore, for the three design points, considering the average RMSE of 0.0073 between the NLS results using actual data from 1 to 30 cycles and the actual data, it is assumed that the extrapolated median results using the NLS method do not significantly differ from the actual EOL. The results presented in Table 6 demonstrate that the EOL_{mid} obtained with the actual data falls within the 95% confidence interval of the EOL estimated by the proposed method.

When comparing the maximum EOL_{mid} point with the initial point, Table 6 shows an increase of approximately 99.17 cycles in EOL_{mid}, based on the estimated EOL_{mid} in

the operational aspect. However, there is a notable decrease of about 72.75% in SEA_{mean}, representing the performance aspect. This trend is also observed at the knee point, where the operational aspect shows an EOL_{mid} increase of approximately 19.66 cycles, but a performance decrease of around 51.32% in SEA_{mean}. These observations highlight the trade-off between SEA_{mean} (performance) and EOL_{mid} (operational aspect, including degradation) in NSH unit cells. The initial design point has a high SEA_{mean} value but a very low EOL_{mid} value in terms of service life, presenting a risk of breakage in case of repeated use. The results of the MOO show that the expected EOL_{mid} result for the SEA value at the initial design point and the corresponding value is 4.72 cycles at t = 3.147 mm h = 2.926 mm, which is an improvement in life and performance compared to the initial design point.

Moreover, analyzing the data from Tables 5 and 6, it can be inferred that if the target life is set to 20 cycles, the knee point emerges as the most reasonable design, considering the estimated EOL_{mid}. Conversely, if reusability is not a priority, the point with the maximum SEA_{mean} value appears to be the optimal design choice. Consequently, this suggests that the most reasonable design can be determined from the Pareto optimum set, depending on the target life set by the designer.

8. CONCLUSION

In this study, a novel design framework for NSH unit cells was proposed, focusing on energy absorption and reusability. SEA_{mean} was considered as a performance metric, while EOL_{mid} estimation relied on operational degradation from cyclic compression. Using the repeated compression test data of 3D-printed NSH unit cell, a trade-off relationship between SEA_{mean} and EOL_{mid} was identified through Pareto frontier analysis employing the NSGA-2 optimization algorithm. From the MOO results, it is evident that establishing a criterion for the target life enables the identification of a viable design point for that lifespan. This approach not only facilitates the lifespan-oriented design of NSH unit cells but also highlights the potential for its application in the design of multi-layer NSHs or similar negative stiffness structures. In the application of these structures, the lifespan of the structure is factored into the design process so that the time to repair or replace the structure can be considered and reflected in the design phase. This framework can be expected to facilitate decision-making based on information about the predicted health of the structure at the design stage and provide possibilities for prognostics and health management (PHM) for design.

Finally, future work aims to develop a PHM framework for robust design that can account for uncertainties or noise that may occur during the manufacturing process and in the testing or operational environment, as efforts continue to predict the health more precisely and EOL of these structures.

ACKNOWLEDGEMENT

This work was supported by the (NRF) grant funded by the Korea government (MIST) (No. 2020R1A5A8018822, 2021R1A2C1013557, and 2022H1D3A2A01052491)

NOMENCLATURE

t	thickness of curved beam
h	central height of curved beam
t_{sc}	scaled thickness of curved beam
h_{sc}	scaled central height of curved beam
Q	t/h
c	height of curved beam for length
l	length of curved beam (= 60 mm)
l_x	horizontal length of curved beam (= 0 ~ 60 mm)
b	width of curved beam (= 12 mm)
ρ	density of PLA/PHA filament (= 1.24 g/cm^3)
F	normalized force
f	reactive force
w	displacement
d_n	normalized displacement (= w/h)
m	mass of curved beam
ξ	parameter of degradation equation
$\boldsymbol{\theta}$	global function's coefficients
p	parameter of correlation function
L	likelihood
m	mass
σ^2	variance
y	observed data
y_{pred}	predicted value
y_{true}	true value
$\hat{y}$	output of Kriging
EA_{SC}	Scaled EA from original data
N	number of observations
N_{NS}	number of negative stiffness occurrences
N_{NNS}	number of non-negative stiffness occurrences
N_d	number of dimensions
N_p	number of parameters
c_{NS}	class for occurrence of negative stiffness
c_{NNS}	class for occurrence of non-negative stiffness
s	standard deviation
$\boldsymbol{\Psi}$	Jacobian matrix
n_p	number of parameters
β	bandwidth
$\mathbf{g}$	bases of global function
$\mathbf{G}$	Matrix of bases of global function
$\mathbf{M}_\xi$	variance of parameters for degradation equation
$\boldsymbol{\Gamma}$	correlation matrix
$\boldsymbol{\gamma}$	correlation vector
$\mathbf{x}$	vector of design variables *t* and *h*

REFERENCES

Chen, S., Wang, B., Zhu, S., Tan, X., Hu, J., Lian, X., Wang, L., & Wu, L. (2020). A novel composite negative stiffness structure for recoverable trapping energy. *Composites Part A: Applied science and*

manufacturing, *129*, 105697. doi: 10.1016/j.compositesa.2019.105697

Correa, D.M., Klatt, T., Cortes, S., Haberman, M., Kovar, D. & Seepersad, C. (2015), Negative stiffness honeycombs for recoverable shock isolation, Rapid Prototyping Journal, Vol. 21 No. 2, pp. 193-200. Doi: 10.1108/RPJ-12-2014-0182

Correa, D.M., Seepersad, C.C. & Haberman, M.R. (2015), Mechanical design of negative stiffness honeycomb materials. *Integr Mater Manuf Innov* **4**, 165–175. doi: 10.1186/s40192-015-0038-8

Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Parallel Problem Solving from Nature PPSN VI: 6th International Conference,* September 18–20, Paris, France. doi:10.1007/3-540-45356-3_83

Debeau, D. A., Seepersad, C. C., & Haberman, M. R. (2018). Impact behavior of negative stiffness honeycomb materials. *Journal of Materials Research*, *33*(3), 290–299. doi:10.1557/jmr.2018.7

Gonabadi, H., Yadav, A. & Bull, S.J. (2020). The effect of processing parameters on the mechanical characteristics of PLA produced by a 3D FFF printer. *Int J Adv Manuf Technol* **111**, 695–709. doi:10.1007/s00170-020-06138-4

Ha, C. S., Lakes, R. S., & Plesha, M. E. (2019). Cubic negative stiffness lattice structure for energy absorption: Numerical and experimental studies. *International Journal of Solids and Structures*, *178*, 127-135. doi: 10.1016/j.ijsolstr.2019.06.024.

Klatt, T. Michael, H. Seepersad, C.C. (2013), Selective Laser Sintering of Negative Stiffness Mesostructures for Recoverable, Nearly-Ideal Shock Isolation, *2013 cInternational SFF Symposium*, August 12-14, Austin, doi:10.26153/tsw/15653

Kim, N. H., An, D., & Choi, J. H. (2017). Prognostics and health management of engineering systems. Switzerland: Springer International Publishing.

Letcher, T, & Waytashek, M. (2014) "Material Property Testing of 3D-Printed Specimen in PLA on an Entry-Level 3D Printer." *Proceedings of the ASME 2014 International Mechanical Engineering Congress and Exposition*. November 14–20, Montreal, Quebec, Canada. doi:10.1115/IMECE2014-39379

Liu, F., Jiang, X., Wang, X., & Wang, L. (2020). Machine learning-based design and optimization of curved beams for multistable structures and metamaterials. *Extreme Mechanics Letters*, *41*, 101002. doi: 10.1016/j.eml.2020.101002

Liu, Y., Jiang, W., Hu, W., Ren, L., Deng, E., Wang, Y., Song, C. & Feng, Q. (2023). Compressive strength and energy absorption characteristics of the negative stiffness honeycomb cell structure. *Materials Today Communications*, *35*, 105498. doi: 10.1016/j.mtcomm.2023.105498

Matthews, J., Klatt, T., Morris, C., Seepersad, C. C., Haberman, M. & Shahan, D. (2016). Hierarchical Design of Negative Stiffness Metamaterials Using a Bayesian Network Classifier. ASME. *J. Mech. Des*. 138(4): 041404. doi:10.1115/1.4032774

Morris, C., Bekker, L., Haberman, M. R. & Seepersad, C. C. (2018). Design Exploration of Reliably Manufacturable Materials and Structures With Applications to Negative Stiffness Metamaterials and Microstereolithography. ASME. *J. Mech. Des*. 140(11): 111415. doi:10.1115/1.4041251

Rasmussen, C. E. & Williams, C. K. (2006). Gaussian processes for machine learning. Cambridge, MA: MIT press.

Shahan, D. W. & Seepersad, C. C. (2012). Bayesian Network Classifiers for Set-Based Collaborative Design. ASME. *J. Mech. Des*. 134(7): 071001. doi:10.1115/1.4006323

Shan, S., Kang, S.H., Raney, J.R., Wang, P., Fang, L., Candido, F., Lewis, J.A. & Bertoldi, K. (2015), Multistable Architected Materials for Trapping Elastic Strain Energy. Adv. Mater., 27: 4296-4301. doi:10.1002/adma.201501708

Scott, D. W. (2015). Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Tan, X., Chen, S., Zhu, S., Wang, B., Xu, P., Yao, K., & Sun, Y. (2019). Reusable metamaterial via inelastic instability for energy absorption. *International Journal of Mechanical Sciences*, *155*, 509-517. doi: 10.1016/j.ijmecsci.2019.02.011

Wang, B., Tan, X., Zhu, S., Chen, S., Yao, K., Xu, P., Wang, L., Wu, H. & Sun, Y. (2019). Cushion performance of cylindrical negative stiffness structures: Analysis and optimization. *Composite Structures*, *227*, 111276. doi: 10.1016/j.compstruct.2019.111276

Xu, H. (2020). Constructing Oscillating Function-Based Covariance Matrix to Allow Negative Correlations in Gaussian Random Field Models for Uncertainty Quantification. ASME. J. Mech. Des. 142(7): 074501. doi:10.1115/1.4046067

Zhakatayev, A., Kappassov, Z., & Varol, H. A. (2020). Analytical modeling and design of negative stiffness honeycombs. *Smart Materials and Structures*, *29*(4), 045024. doi:10.1088/1361-665X/ab773a

Zouaoui, M., Gardan, J., Lafon, P., Makke, A., Labergere, C., & Recho, N. (2021). A finite element method to predict the mechanical behavior of a pre-structured material manufactured by fused filament fabrication in 3D printing. *Applied Sciences*, *11*(11), 5075. doi:10.3390/app11115075

BIOGRAPHIES

Hyung-Do Kim is currently an integrated MS and Ph.D student at Pusan National University, Busan, South Korea. He received his B.S. degree in department of mechanical engineering from Kyonggi University, Suwon, South Korea, in 2022. His research interests include data driven design and design for manufacturing.

Taemin Noh is currently a MS student at Pusan National University, Busan, South Korea. He received his B.S. degree in department of mechanical engineering from Pusan National University, Busan, South Korea, in 2022. His research interests include data driven design and design for manufacturing.

Young-Jin Kang received the Ph.D. degree in mechanical engineering from Pusan National University, Busan, South Korea. He has been working as a postdoctoral researcher in Research Institute of Mechanical Technology, Pusan National University, Korea. His research area is uncertainty quantification, design under uncertainties, data driven design, and fault detection and diagnosis.

Nam-Ho Kim received his Ph.D. degree in mechanical engineering from University of Iowa, United States. He has been working as a professor in the Department of Mechanical engineering at University of Florida, Gainesville, United States. His research areas include design under uncertainty, prognostics and health management, uncertainty quantification, and nonlinear structural mechanics.

Yoojeong Noh received the Ph.D. degree in mechanical engineering from University of Iowa, United States. She has been working as an associate professor in the school of mechanical engineering in Pusan National University, Busan, South Korea. Her research interests include computational mechanics, design under uncertainties, data driven design, and fault detection and diagnosis.

Mastering Training Data Generation for AI - Integrating High-Fidelity Component Models with Standard Flight Simulator Software

Andreas Löhr[1], Conor Haines[2]

[1,2]*Linova Software GmbH, München, Bavaria, 80805, Germany*
andreas.loehr@linova.de
conor.haines@linova.de

ABSTRACT

The German state-funded aviation research project "Real-time Analytics and Prognostic Health Management" (RTAPHM) envisioned fully automated urban air services executed by autonomous drones and infrastructure controlled by a digital system. Research was focused on utilizing onboard real-time diagnostics to enable AI-driven UAV capability predictions. These predictions increased the reliability of upfront service commitments. The use case selected to demonstrate these elements was organ transport. The project delivered an end-to-end demonstrator incorporating a virtual fleet of drones with onboard diagnostics to provide data for the platform decision logic.

The project followed a „digital-twin-first" approach to overcome a common bootstrapping problem faced by data-driven applications. That is, the lack of in-service data for exploration, prototyping and training of diagnostic and prognostic approaches during the concept and early development phases. Due to the upfront development of physical high-fidelity simulation models for the monitored components, a digital twin – of the portion of the twin that resembles the physical behavior – was used to generate data and facilitate preliminary exploration, prototyping and training. Digital twins were further employed to allow evaluation of what-if scenarios and identify the optimal future operation parameters of a drone.

Development of the RTAPHM digital twin involved a multi-disciplinary team of members distributed across different organizations and locations. Successful realization of the digital twin depended on early integration testing, performed in high frequencies, which generated continuous feedback regarding technical and conceptual issues. Within the research project we developed MOLE, an engineering tool for automating the integration of distinct simulation components, into a single system simulation driven by commercially available flight simulator software. Here, we showcase the internal mechanisms of the tool and demonstrate its abilities to generate a Docker-based executable for efficient data generation in the cloud. We also show our approach to online visualization, fault insertion, batch integration testing and debugging the digital twin executable. We also report on the utilization of MOLE in assembling the final RTAPHM demonstrator (Löhr, 2023).

1. OUTLINE

The document first introduces the RTAPHM project with a focus on the use and purpose of digital twins. This leads to our primary project contribution: MOLE, a software tool assisting in the fast integration of digital twin components. After describing the core principles of MOLE, we report from our experience in using MOLE to build the digital twin for the RTAPHM demonstrator. We conclude with a suggestion for areas of future work.

2. INTRODUCTION

The project Real-time Analytics and Prognostic Health Management (RTAPHM) was a joint government-funded research endeavor in Germany It was in the aviation domain and included SMEs, academia, and industry as partners. The project concluded in June 2023. We participated as an SME partner focused on simulation and software engineering IVHM systems.

The project contributed to the field of urban air services, such as person and cargo transport, imaging services, etc. The vision of a fully automated urban air services platform drove the work. Here, customers could book a variety of different services via low-threshold access channels, e.g., via their smart phone. Once booked, the service would be carried out by autonomous unmanned air vehicles (UAVs) supported by autonomous infrastructure, e.g., for mounting payload or loading cargo. The platform would have to work digitally to

benefit from automation. AI technologies would support the orchestration multiple booked services in parallel and the handling of problems which could not be sufficiently formalized upfront.

Obviously, such a bold vision cannot be accomplished by a single research project. Therefore, project research focus was narrowed to a single use case highlighting a specific problem within that use case. The project selected organ transport as that use case. By the means of a digital service platform, the personnel of a donor hospital would book (air) transport to a specific receiver hospital, as depicted in Figure 1. The last mile between the hospitals and the next logistics hub should be covered with transport UAVs.

Figure 1. Synopsis of selected use case.

Within the use case, the project focused on the issue of how to make reliable service commitments given a fleet of UAVs (or other transportation vehicles for long haul routes) depending on their current and future health state. That is, build a platform capability to assess whether a specific service request can be completed with the available resources. This question should be answered by a comprehensive situational awareness picture of the fleet's conditions, mainly driven by real-time diagnostics and prognostic capabilities, integrated into the onboard data processing of the UAVs. Figure 2 depicts the research focus on a high level: for selected components of an artificial UAV (fuel cell, servo and pusher motor) a monitoring concept should be established and implemented. The obtained (gradient of the) health data should be used to make reliable maintenance predictions for the components. Finally, by having predicted the future maintenance burden – thus, the availability of individual UAVs – incoming service bookings for organ transports would be committed reliably within a timeframe where the selected UAV is available and maintenance free.

3. RTAPHM DIGITAL TWIN

The project selected three components to be monitored: servo motors, a fuel cell and pusher propeller system. For all components, a monitoring concept, including diagnostics and prognostics, was developed, and included in the laboratory demonstrator.

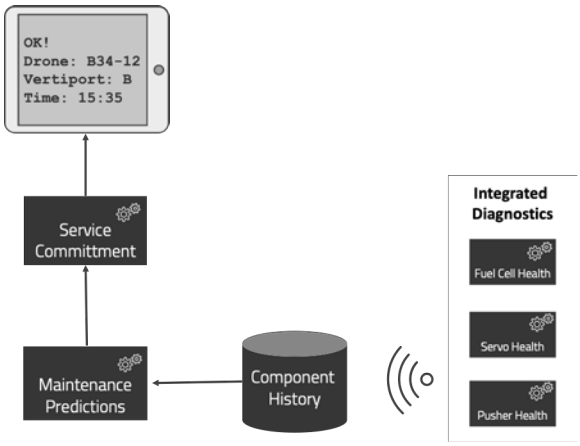


Figure 2. Research focus.

3.1. Motivation

The project faced a common bootstrapping problem, as depicted in Figure 3: incomplete, insufficient, or total absence of operational data from the components in question (the specific configuration of the targeted UAV, as well as the reason for the data absence, is out of the scope of this writing). There was no foundation for conducting exploration or conceptual studies.

Figure 3. Problem statement.

3.2. Digital-Twin-First Approach

The concept of a digital twin is known in both academia and industry. A digital twin is typically derived by observing (data from) an existing system using specific frameworks and methodologies, as for instance Vuckovic, Prakash, and Burke (2023) have shown. The project decided to overcome the bootstrapping problem by employing simulation. The project already incorporated the usage of digital twins within the digital platform. This existing capability and the absence of existing physical components informed the decision to proceed with a "digital-twin-first" approach instead of creating the digital twin afterwards.

To accomplish this, the partners agreed to provide validated high-fidelity component simulation models adapted from previous undertakings. These models formed the bases of the

digital twin. Within the project, these models were adapted according to the performance requirements of the targeted UAV platform and equipped with additional technical features to facilitate integration. Additionally, to stimulate the models with realistic input data, the project decided to use a commercially available flight simulator software.

The solution approach is depicted in Figure 4: a commercial flight simulator should be equipped with a plugin, to extract and send specific data streams (e.g., electrical power demand, thrust demand, environmental data, etc.) to the simulation models as input. Optionally, the models should be operated in a closed loop with the flight simulator, depending on its capabilities. Being stimulated in a realistic way, the simulation models should provide simulated sensor data for the health assessment algorithms. Also, as Darrah, Frank, Quinones-Grueiro, and Biswas (2021) have pointed out, simulation allows generation of run-to-failure data – which would be an unsafe undertaking for a real system.

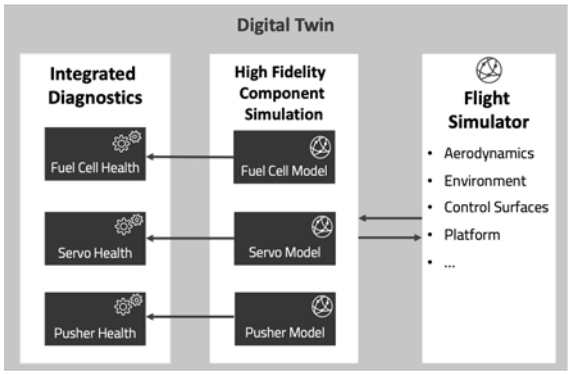


Figure 4. Solution approach.

3.3. Conceptual Challenges

Performing digital-twin-first imposes certain challenges to the modeling process and the overall concept.

This approach is used to obtain data from something that does not yet exist. But how can something that does not yet exist, be modeled? The project addressed this challenge by adapting existing work and narrowed the challenge down to finding the right scaling for the target environment.

There is a further challenge of modeling the right inputs and output of each simulation component. This was addressed by employing agile development techniques, such as a high frequency of iterations and many integration attempts with close feedback amongst all partners.

Finally, the challenge of representativeness and validity remains. Additionally, one must answer the question of how credible the resulting diagnostic and prognostic approaches can be, if they are developed on pure simulated data. There is a risk that simulation development is driven to produce what the exploiting modules expect and vice versa rather than reflecting a realistic and useful abstraction of the potential platform. This challenge was addressed by using (adapting) simulation models which had been created independently and validated in isolation. However, using simulations in this context can only be a first measure to parallelize development. As soon as the first set of "real" operational data becomes available, it must be used to tune the simulation.

3.4. Integration Challenges

Our task in the RTAPHM project, amongst others, was the provision of an integrated executable digital twin to be used by the partners to generate data according to their scenarios. Along with the task itself came the necessity to not only perform the physical integration just once towards the end of the project, as it would have been in a waterfall organized project. Instead, to adhere to the agile mindset of the digital twin development process, we had to be able to perform the physical integration as often as possible. We identified a set of challenges in the context of a multi-organization project that we had to address.

- in what form should a model be delivered?
- how can the intellectual property of a component providing partner be protected?
- what needs to be provided so that a specific model can be integrated with others?
- how can the interaction of two models be tested?
- how to be test the nominal and erroneous behavior?
- how can a quick turnaround be performed

4. MOLE

While the conceptual challenges were addressed by our partners, we worked on providing a solution to the integration challenges that the project was facing. We provided that solution on the form of MOLE – a desktop software tool for automating large portions of the digital twin integration process. Besides the capability to automatically create an executable digital twin from the provided simulation models, the tool provided support for the actual development cycle, and could be used directly by the partners.

4.1. Model Format

Our partners agreed to deliver the models as C code. Some chose to generate the C code using graphical modeling tools, such as Matlab/Simulink, while others provided custom code. Whatever the source, all models to adhered to a common interface concept, which was derived from the way that Matlab exports models. It consisted of:

- structures representing the internal state of a model
- an initialization function for the structures
- a stepping function, accepting (pointers to) the structures

- a custom naming scheme for parameters to establish semantic consistence

Intellectual property was protected by giving the partners the option to provide their models in pre-compiled object code, accompanied with their respective header files. Partners whose models depended on the models of other partners agreed on the specific information to be exchanged, and then specified the required technical parameters. We acted as a central parameter registry to enforce consistency.

As part of the conventions, simulated fault behavior was triggered via a set of up to four parameters. This set consisted of one input parameter acting as a plain on/off switch, a second defining the degree (severity) of the simulated fault, a third setting the point in time at which the fault should become effective, and a fourth setting whether the fault should become effective immediately, or if a degradation towards the specified severity should be simulated.

4.2. Automated Wiring

A significant portion of the integration work consists in the correct wiring of a model's outputs to one or mode inputs of dependent models, according to a specification given by the model creators. By "wiring" in the technical sense, we mean the temporal storage of a model's output in memory, so that it can be read by all dependent models once they start calculating their next cycle. Figure 5 shows an example of MOLEs mapping browser, a visualization of all detected inputs of a specific model, and the assigned outputs for each inputs based on naming conventions.

Model Variables	Type	Source	Initial Value	Internal Resolution
EMA_CBS	Simulink			1.0E-5
Inputs				
EMA_DBBs_PsBll[120]	double			
EMA_DNt_PsNtl[12]	double	EMA_DNt.EMA_DNt_PsNtl[12]		
EMA_DSc_PsScl[12]	double	EMA_DSc.EMA_DSc_PsScl[12]		
EMA_TRM_TNt[9]	double	EMA_TRM.EMA_TRM_TNt[9]		
EMA_TRM_TSc[9]	double	EMA_TRM.EMA_TRM_TSc[9]		
SIMCONTROL_DMG	double		1	
Outputs				
EMA_CGB	Simulink			1.0E-5
EMA_CPB	Simulink			1.0E-5

Figure 5. Component I/O and mapping browser.

To relieve the burden of creating specification documents and harmonizing those documents among the partners, we exploited the project wide naming conventions for model parameters (inputs/outputs). Having established the conventions, we programmed MOLE to use them in inspecting (parsing) each provided simulation module and generating "glue" code for correct parameter exchange. To cater for exceptional cases, we incorporated the ability to manually override the automated wiring.

Figure 6 aims at illustrating what we mean with "wiring". The figure depicts a simulation of two aircraft systems, the hydraulic system and the fuel system. Each system simulation is decomposed into smaller simulation blocks, such as pumps, circuitry, and actuators. Each block exposes specific inputs (upper ports of each block) and outputs (lower blocks), whereas each output models either a sensor (e.g., a pressure sensor) or a specific physical property (e.g., a specific pressure) that provides input for another simulation block (e.g., the pressure output of the pump acts as the input of various actuators). A specific interface block models the interdependency between the two systems. Finally, some of the inputs will be fed from an external aircraft simulation. MOLE is able to generate code (glue code) for the data exchange, even for high exchange frequencies (see section 4.3).

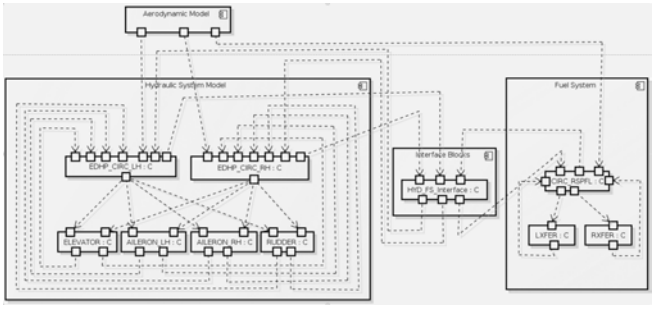


Figure 6. Glue code for automated wiring.

4.3. Parallel Computing

Due to their intended use as source for health assessment algorithm training data, the models performed complex calculations demanding high CPU power. The necessity for small temporal solutions, in the magnitude of 10^{-4}s – 10^{-6}s, drove the need for computational resources. Additionally, different models exhibited different temporal solutions, which were in general not multiples of each other. To benefit from multi-core processors, we introduced a concept for executing each model block in isolation with constant inputs for a maximum number of steps before the model block becomes unstable, and then halting the execution while each model block was updated with external inputs according to the specified wiring (while respecting overrides by expression). The concept was based on two frequencies. The "internal I/O frequency" determined the rate by which the model block execution was halted for the sake of parameter updates. As this lower frequency is reduced, overall execution time is also reduced, as a smaller relative fraction of time was used for parameter exchange and threading overhead. The "external I/O frequency" determined the rate in which supporting functions like graph plotting, data recording or other custom plugins were triggered. Based on that concept, MOLE supported setting a constant number of utilized CPU cores or dynamically adjust this number based on runtime performance analysis.

Figure 7 depicts the main parallel execution loop. First, each simulation block is iterated in its own thread, with constant input values for a block-individual maximum number of iterations that leaves the block numerically stable. Once each block is finished, the respective block outputs are copied to

the inputs according to the wiring. Then, further supporting functions are executed (typically on the current outputs), see sections 4.4 and 4.5. Then the loop repeats.

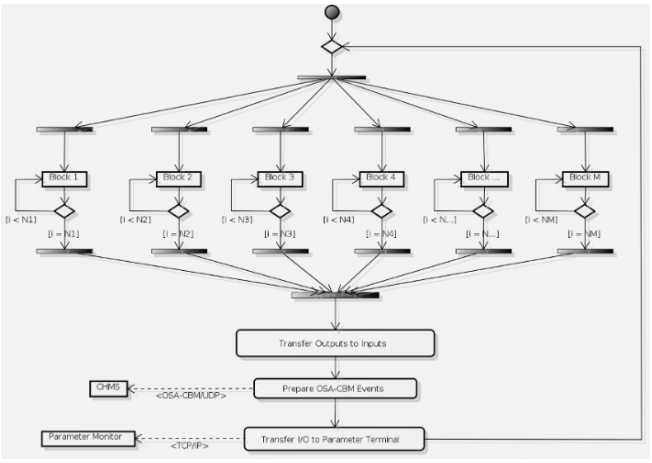


Figure 7. Main execution loop.

4.4. OSA-CBM Sampling

Based on our experience with implementing MIMOSA standards (Löhr & Buderath, 2014) we encouraged the project to follow the OSA-CBM design principles for the data processing chain. Consequently, the integrated diagnostics component of the digital twin was equipped with an OSA-CBM compliant sensor data interface, which was implemented using our proposed implementation of the binary OSA-CBM messaging protocol described by Drever, Naughton, Nagel, Löhr and Buderath (2016), which also discuss challenges and opportunities of MIMOSA standards in general. To drive the integrated diagnostics with data from the simulation, we enhanced MOLE with an OSA-CBM-compliant sampling function that was tied to the external I/O frequency. Selected model outputs, specifically those that simulated sensors, were sampled with the frequency required by the defined monitoring concept in the diagnostics layer. Then, the samples were automatically wrapped into OSA-CBM DM DataSeq events and transmitted to the integrated diagnostics via UDP.

4.5. Supporting Features

The core features of MOLE kept the inputs and outputs consistent, and ensured all models could be compiled. This allowed for quick iterations. Further, we added the following functions to enhance the testing process:

- Stepping: running a single model or a set of models, and halting/resuming the execution on demand, to inspect the current parameter sets
- Recording: marking a set of model inputs and outputs to be written to a CSV file for offline inspection, or for exchange within the project
- Fault Insertion: picking up on the naming conventions for faults, the fault input quadruples were recognized, and presented to the developer in a graphical user interface for interactive control
- Visualization: plotting the development of selected parameters over time to visually inspect the behavior of the model. An example is given in Figure 8.
- Parameter Override: we added an online compiler for simple expressions bound to model parameters. If set, the result of the expression was set (input parameter) or distributed (output parameter), resulting in a dynamic override of the original model wiring.
- Scripting: we added a simple scripting feature which allowed the association of model executions with time-indexed parameter overrides for that execution. By supporting an operation mode in which MOLE executes all scripts in a certain directory tree, users could generate data for different constellations, thus, having a set of repeatable test cases at their disposal.

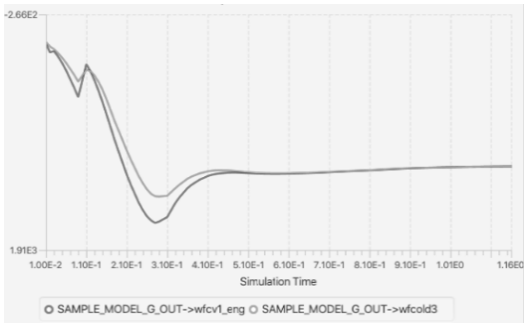


Figure 8. Integrated visualization.

5. VALIDATION

Validation of MOLE functionality was accomplished by its use in supporting the projects continuous integration process and building of the final deliverables. The constituents of the RTAPHM digital twin were:

- model of a pusher motor in several blocks
- model of a fuel cell in several blocks
- model of a servo motor in several blocks
- diagnostic module for pusher motor
- diagnostic module for fuel cell
- diagnostic module for pusher motor
- prognostic module for fuel cell
- interface module for pushing telemetry data to digital platform

- interface module for exchanging data with COTS flight simulator software
- bundled flight simulator module consisting of Gazebo, PX4, and a communication adapter towards the RTAPHM interface module

As validation evidence and as deliverables to the RTAPHM project, we created two artifacts.

First, a MOLE-based integration process for the digital twin. Here, Gazebo/PX4 was used as the "aerodynamics driver" to stimulate three high-fidelity physical simulation models. These models, in turn were used to generate sample and training data for three individual implementations of monitoring concepts.

Second, an interactive digital twin was created as a result of the integration process. This was used during project demonstration runs to create a live feed of (simulated) sensor data into the running diagnostic and prognostic modules, to show the interactive reaction to different fault scenarios.

Figure 9 depicts the result: simulation models for fuel cell, servo and pusher motor formed the RTAPHM system simulation. Some of the system simulation's inputs were fed with data from a customized fixed wing VTOL-capable UAV of which the physical and aerodynamic properties were simulated using Gazebo. Both system simulation and Gazebo were step-synced. PX4 was used to auto-pilot the UAV along a set of waypoints.

Figure 9. RTAPHM virtual aircraft.

6. Future Work

The validation work points out paths for future work. The most significant findings, which we are currently pursuing, are presented below.

- ***support further modelling tools***: to enlarge the possible target audience for MOLE users, the support of further modelling tools, other than just Matlab/Simulink, seems promising.
- ***support open standards***: another aspect of widening the target audience of MOLE emerges from the support of open interface standards, such as the Functional Mockup Interface (FMI), instead of relying on parsing proprietary coding patterns.
- ***integrate with Asset Administration Shell***: the current efforts around (AAS) focus on standardizing the digital representation of assets. MOLE could be integrated with AAS compliant repositories to allow users to choose from pre-build simulation components.
- ***cloud-native technology***: currently, MOLE is designed as a desktop-based tool, and thus is limited by the resources of the computing platform it runs on. By extracting the headless MOLE computing core and basing in on state-of-the-art cloud-native technology, these limitations can be overcome, and data stakeholders can scale the required resources on demand.
- ***service platform***: finally, combining all the previously mentioned streams of future work, we envision a cloud-based service platform, where stakeholders can upload own contributions, and build complex simulations from own and 3rd party simulation components, depending on their specific needs for training) data, and – by employing AAS and FMI concepts – without having to expose their intellectual simulation property to the public.

Using the MOLE digital twin, the project was able to showcase a core use case for automated functional dependency analysis: the injection of a fault in the cooling system's filter (clogging) caused the relevant areas to heat up, causing a decrease in the efficiency of the fuel cell, in turn, decreasing the remaining useful life of the fuel cell.

7. Conclusion

We presented MOLE, a tool for supporting the automation of the integration of software-based system simulations. The tool facilitates short integration cycles for agile project setups and provides specific debug and testing features. We showed the benefits of a structured and automated integration process for distributed research projects with different interests of individual partners.

We reported on the benefits of MOLE for rapid assembling of demonstrators for PHM research projects (though MOLE is not limited to that application domain). MOLE was used to support the integration process of the RTAPHM digital twin, consisting of commercial flight simulator software driving high-fidelity models of aircraft components.

We also showed that the bootstrapping and early development phases of data-driven application projects can benefit from artificially generated data to overcome the lack of initial data for first exploration and training. In particular, the data and communication architecture of the envisioned system can be explored regardless of the origin of the utilized data.

We finally recommend that OSA-CBM design principles and communication standards should not only be adopted for productive modules of a data processing chain. Designing data generators as OSA-CBM compliant data sources (see 4.4) facilitates integration testing, as the productive interface of the modules can be transparently stimulated.

ACKNOWLEDGEMENT

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

This work was supported by the ministry of commerce and climate protection (Bundesministerium für Wirtschaft und Klimaschutz) based on a resolution of the German Government (Deutscher Bundestag).

REFERENCES

Asset Administration Shell (AAS), https://www.iec.ch/dyn/www/f?p=103:38:607572709001913::::FSP_ORG_ID,FSP_APEX_PAGE,FSP_PROJECT_ID:1250,23,103536, IEC, International Electrotechnical Commission,

Darrah, T., Frank, J., Quinones-Grueiro, M., & Biswas, G. (2021). *A Data Management Framework & UAV Simulation Testbed for the Study of System-level Prognostics Technologies*. Annual Conference of the PHM Society, 13(1). https://doi.org/10.36001/phmconf.2021.v13i1.3030

Drever, J., Naughton, H., Nagel, M., Löhr, A., & Buderath, M. (2016). *Implementing MIMOSA Standards*. PHM Society European Conference, 3(1). https://doi.org/10.36001/phme.2016.v3i1.1647,

Löhr, A. (2023). *Realisierung einer auf operationellen Daten basierenden Plattform zur Qualitätskontrolle und zur Fortschrittsüberwachung bei der Erbringung von digitalen oder digital gestützten Flottendiensten*, *Technische Informationsbibliothek (TIB)*, June 2023

Löhr, A., & Buderath, M. (2014). *Evolving the Data Management Backbone: Binary OSA-CBM and Code Generation for OSA-EAI.* PHM Society European Conference, 2(1). https://doi.org/10.36001/phme.2014.v2i1.1487,

Functional Mockup Interface (FMI), https://fmi-standard.org, Modelica Association c/o PELAB, IDA, Linköpings Universitet S-58183 Linköping Sweden

Vuckovic, K., Prakash, S., & Burke, B. (2023). *A Framework for Rapid Prototyping of PHM Analytics for Complex Systems using a Supervised Data-Driven Approach*. Annual Conference of the PHM Society, 15(1). https://doi.org/10.36001/phmconf.2023.v15i1.3480

BIOGRAPHIES

Andreas Löhr received his M.Sc. degree in Computer Science from the Technical University of Munich in 2001 (Informatics, Diplom) and earned his doctoral degree in Computer Science from Technical University of Munich in 2006. For 6 years he worked as a software engineer at Inmedius Europa GmbH in interactive technical publications and researched in the field of wearable computing. He co-founded Linova Software GmbH in 2008 as managing director and worked as coding architect and consultant in both industry and academia projects in the field of aircraft maintenance and fleet optimization. Next to growing his company and concentrating on its strategic development, he specializes in aircraft MRO topics, with a research focus on condition-based maintenance, digital twins and data architectures.

Conor Haines received his B.Sc. degree in Aerospace Engineering from Virginia Polytechnic Institute and State University in 2003 and his M.Sc. degree in Computational Science from the Technical University of Munich in 2011. For 3 years Conor was a test engineer supporting the NASA Near Earth Network, providing simulation support used to guide system development. At his current post with Linova Software GmbH, he is focused on designing and developing embedded IVHM/ISHM architectures, along with a special interest in the application of digital twin simulation frameworks for the generation of early training data for data-driven application, such as enhanced embedded diagnostic and prognostic capabilities.

Model-Based Loads Observer Approach for Landing Gear Remaining Useful Life Prediction

Jonathan Jobmann[1] and Frank Thielecke[2]

[1,2] *Institute of Aircraft Systems Engineering – Hamburg University of Technology, Hamburg, 21129, Germany*
jonathan.jobmann@tuhh.de
frank.thielecke@tuhh.de

ABSTRACT

Implementing health monitoring methods for aircraft landing gears holds the potential to prevent premature component replacements and optimize maintenance scheduling. Therefore, this paper introduces a fundamental framework for fatigue monitoring and subsequent steps for predicting the remaining useful life of landing gears. A key component of this framework is the model-based load observer, which lays the groundwork for subsequent remaining useful life prediction steps. This load observer will be analysed in detail in this paper. The model-based approach is specifically designed for observing the loads on civil aircraft landing gears during touchdown, utilizing signals from in-service sensors. To evaluate the load observation method, a flexible multibody simulation model is introduced to generate synthetic data sets of aircraft in-service data and the corresponding landing gear loads, given the unavailability of real in-service and recorded landing gear load data. The load observation method is applied to synthetic in-service data across various virtually performed landing scenarios, offering a proof of concept along with extensive analysis of parameter uncertainties and additional factors influencing observation quality. Through this analysis, certain challenges to the observation method are identified that require further investigation in subsequent research efforts.

1. INTRODUCTION

Optimizing aircraft life cycle management significantly contributes to enhancing profitability and maintaining competitiveness within the aircraft industry, while also facilitating the achievement of ambitious climate objectives. One essential aspect of an aircraft's life cycle involves its operational life, including maintenance. Emerging maintenance strategies, such as condition-based, predictive, and prescriptive maintenance, prioritize health-oriented approaches aiming to optimize aircraft operating life by enhancing performance and safety. The advancement of processes and methodologies for these strategies is facilitated by the growing digitization and improved IT infrastructure, notably through digital platforms that address aircraft operational life and maintenance needs. This technological advancement enables the intensive computation and memory utilization required for certain health monitoring methods, which contribute to an optimized aircraft life cycle management. Especially structural components, such as aircraft landing gears (LG), offer high potential for the meaningful implementation of health monitoring methods.

LG systems must endure a variety of severe loads across different loading conditions. To ensure the structural integrity of the LG, with no detectable fatigue cracking throughout its operational lifespan, the safe life design philosophy is commonly employed in structural LG design (Schmidt, 2021). In this context, the safe life denotes the duration during which the components can operate without experiencing fatigue cracking. At the latest, when this point in time is reached or exceeded, the components are retired from service. Designing with the safe life philosophy entails incorporating scatter factors and estimating fatigue load spectra (SAE International, 2020), often resulting in underestimated individual LG lifespans.

However, by gaining detailed insights into actual loads and fatigue experienced during service, the assessment of LG condition and the prognosis of remaining useful life (RUL) can be performed. Consequently, implementing Structural Health Monitoring (SHM) for LGs through fatigue monitoring methods may help avoid premature replacements and optimize maintenance scheduling. Furthermore, the comprehensive understanding of the actual loads experienced in service opens up opportunities to improve future LG designs (Schmidt & Sartor, 2009).

In recent years, several fatigue monitoring approaches have been developed for aircraft, with some specifically tailored for LGs. One common feature among many of these aircraft fatigue monitoring approaches is the observation of loads prior to fatigue calculation. In (Boller & Buderath, 2007), (Boller & Staszewski, 2004), (Buderath & Neumair, 2007), (Buderath,

2009) and (Schmidt & Sartor, 2009) aircraft load observation approaches for fatigue monitoring are described.

One load monitoring approach involves using flight parameters monitored on the aircraft. Another SHM strategy utilizes additional sensors implemented on the aircraft, such as strain gauges. The data collected from these sensors can then be input into a ground-based digital loads model based on the finite element method. A similar approach, employing strain gauges, is addressed in (Chabod, 2022).

However, this approach incurs additional costs as it requires the installation of extra sensors on each LG to be monitored. Moreover, it increases the risk of sensor failures due to these additional sensors, which could affect aircraft availability or compromise the reliability of the implemented SHM methods. Conversely, employing model-based methods reliant on flight parameters extracted from sources like quick access recorder presents a promising strategy for observing LG loads and fatigue and can effectively mitigate these drawbacks. This approach is also already utilized for detecting transient overloads in LGs (Schmidt & Sartor, 2009).

Explicit research on fatigue monitoring of LGs primarily focuses on the utilization of machine learning methods. In (El Mir & Perinpanayagam, 2021), a machine learning model was proposed to determine load histories of the LG based on sensors onboard the aircraft. Additionally, (Holmes et al., 2016) presented results of a machine learning model calculating LG loads on different runway surfaces using sensor data collected from sensors attached to the LG. However, this approach has the disadvantage of requiring additional sensors to be installed on the LGs.

Addressing this limitation, (Jeong, Lee, Ham, Kim, & Cho, 2020) utilized a landing simulation model to generate synthetic flight parameters and related synthetic LG loads and strains for training machine learning models. Nonetheless, model-based methods offer several advantages over black box models like machine learning models. On the one hand, they are typically more robust and interpretable which is a great advantage within the certification process of aircraft systems. On the other hand, model-based approaches can be more efficient in using data, particularly in scenarios with limited data availability, as they often incorporate prior knowledge about the problem domain.

Therefore, this paper presents a model-based loads observer approach designed specifically for monitoring civil aircraft LG loads without the need for additional sensors, primarily utilizing in-service sensor signals as a foundational element. As the load observation of LG operations is very extensive and comes with various challenges, this paper aims to focus solely on the first landing impact of the main LGs. The developed method constitutes a key component of a comprehensive framework for fatigue monitoring in LGs. This framework, along with the steps for remaining useful life (RUL) prediction, is fundamentally introduced. The overall approach aims to lay the foundation for LG lifecycle management optimization through effective fatigue monitoring and prediction in future work.

The paper is organized as follows. The fatigue monitoring framework and subsequent steps for RUL calculation are introduced in Section 2. For detailed analysis of the model-based loads observer as a key component of the LG fatigue monitoring framework, Section 3 outlines the simulation model utilized for generating synthetic data. This synthetic data is essential for evaluating the loads observation method. Section 4 presents the description and analysis of this method. Finally, the paper concludes with Section 5, which provides a summary and a brief outlook.

2. FATIGUE MONITORING AND PREDICTION

In the following section the LG fatigue monitoring framework and an approach for downstream RUL prediction is presented.

2.1. Fatigue Monitoring Framework

There have been numerous publications addressing fatigue monitoring of aircraft structures, such as (Boller & Staszewski, 2004), (Buderath, 2009), (Dziendzikowski et al., 2021), (JIAO, HE, & LI, 2018), and (Stolz & Neumair, 2008). Additionally, publications by (El Mir & Perinpanayagam, 2021) and (El Mir & Perinpanayagam, 2022) have focused specifically on fatigue monitoring of LG systems. What most of these publications have in common is the proposed application of the Miner rule for calculating a health index of the structures. This rule calculates the cumulated damage D of structures over their life cycle using the equation

$$D = \sum \frac{n_i}{N_i}. \tag{1}$$

Applying n_i cycles with a certain stress amplitude i and the corresponding fatigue life endurance N_i on a structural component is equivalent to the consumption of n_i/N_i of fatigue resistance (Schijve, 2009). When the cumulated damage D reaches 1, failure is expected. Given that the Miner rule is presently utilized in the safe life fatigue analysis for LG certification processes (El Mir & Perinpanayagam, 2022), its application in the fatigue monitoring process of the LGs is evident. Therefore, the proposed fatigue monitoring framework in this paper also relies on the damage calculation using the Miner rule as a central element. By utilizing the Miner rule, many aspects of the fatigue monitoring framework are implicitly defined.

The LG fatigue monitoring framework, as illustrated in Figure 1, is based on the remaining life calculation scheme outlined in (Tinga, 2010) and the steps for safe-life analysis presented in (El Mir & Perinpanayagam, 2022). The objective of the

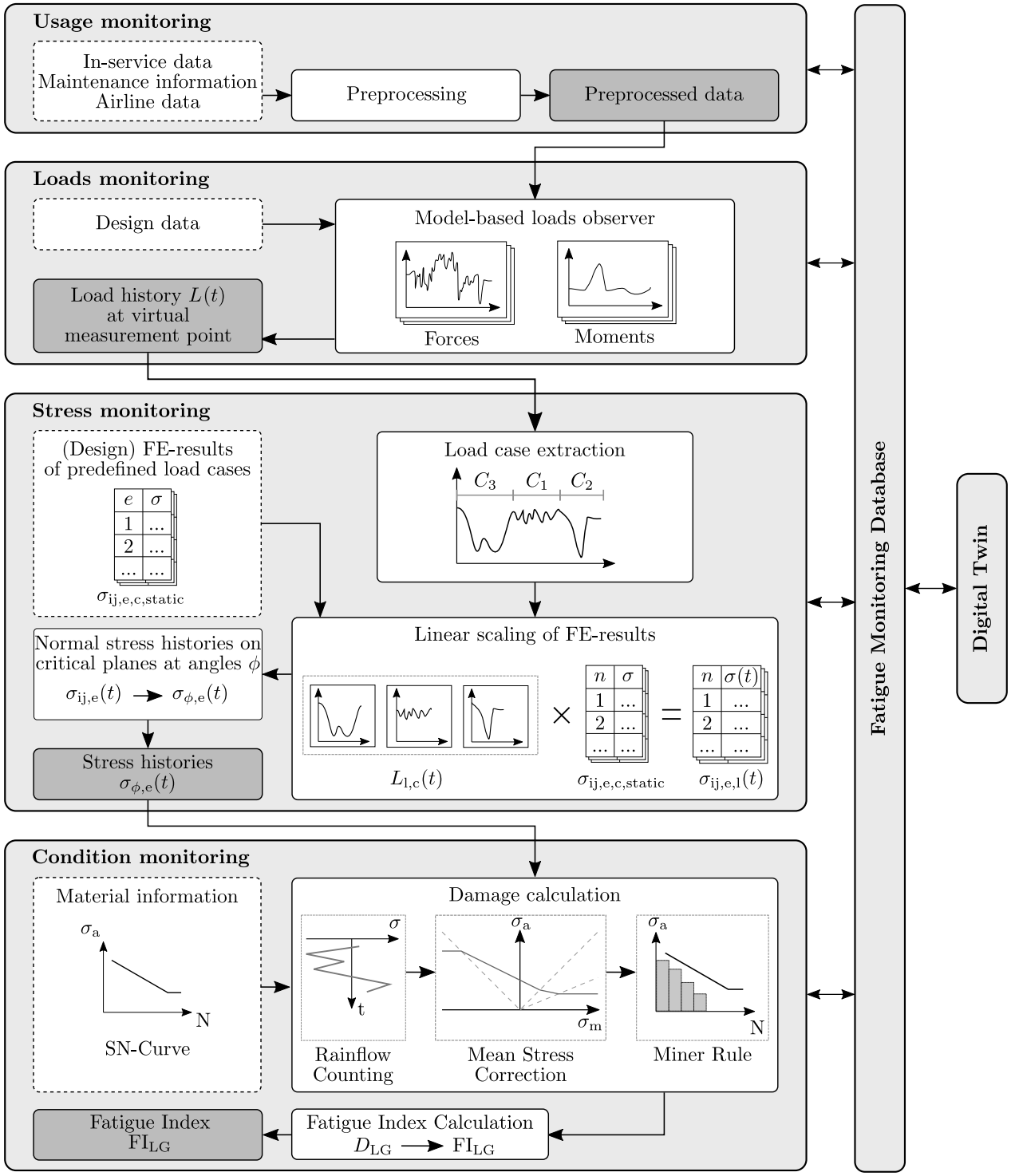


Figure 1. Schematic presentation of the LG fatigue monitoring framework

outlined framework is to present a comprehensive monitoring process for LGs, spanning from raw in-service data recordings over observing LG loads to monitoring LG fatigue and integrating digital twin technology. The framework serves as a basis for subsequent RUL calculation. Since real-time onboard aircraft fatigue monitoring is unnecessary and requires considerable storage and computing capacity, the framework operates offboard. Initially, usage monitoring is conducted within the framework, entailing the recording and storage of essential data, primarily in-service data. This data undergoes

preprocessing, including data cleansing and noise filtering. Subsequently, the preprocessed data is utilized to observe loads on the LG using simulation models constructed with LG design data, thereby virtually emulating the actual LG dynamics (model-based loads observer). The simulation model generates load histories at virtual load measurement points of the LG geometry. These specific locations are also used as load application points in finite element (FE) models for structural analysis. Thus, load histories and FE-models can be combined for subsequent stress monitoring.

To monitor LG structural stress based on LG load histories, stress tensors $\sigma_{\text{ij,e,c,static}}$ are calculated for selected 'hot spots' or across all finite elements e for different load cases c (e.g. specific steering, braking or landing conditions) using static FE design calculations, as depicted in Figure 1. The index ij represents the respective matrix entry of a stress tensor. To associate these stress tensors with observed loads, the load histories from the model-based loads observers are analysed, and specific load cases are extracted. The load histories are segmented into load events $L_{\text{l,c}}(t)$, where the additional index l denotes the index of the load event within the overall load history, and c links the load event to a specific FE load case. The stress tensors are then linearly scaled based on the load histories for each specific load event, resulting in a stress tensor history $\sigma_{\text{ij,e,l}}(t)$ for each finite element e and each load event l. This scaling is achieved through linear superposition by multiplying each load event with its corresponding stress tensor:

$$L_{\text{l,c}}(t) \cdot \sigma_{\text{ij,e,c,static}} = \sigma_{\text{ij,e,l}}(t). \tag{2}$$

It is important to note that each load event $L_{\text{l,c}}(t)$ is characterized by three force time series and three moment time series along the principal axes. However, the stress tensor is scaled by only one time series, which is selected based on the predominant force specific to the load case. Therefore, for each load event $L_{\text{l,c}}(t)$, the load case-specific predominant load is identified and used for scaling.

Afterwards, the stress tensor histories for all load events are chronologically ordered and concatenated for specific finite elements e, resulting in the combined stress tensor histories $\sigma_{\text{ij,e}}(t)$. To ensure accurate fatigue monitoring under complex, multiaxial loading conditions, the critical plane method is employed (Lee & Barkey, 2012). This method assesses stress across various potential planes to identify those where stresses and strains are most likely to cause damage. The stresses $\sigma_{\phi,\text{e}}$ on various planes of finite element e, oriented at angles ϕ under biaxial stress, are calculated using the formula:

$$\begin{aligned}\sigma_{\phi,\text{e}} =& \frac{\sigma_{\text{xx,e}} + \sigma_{\text{yy,e}}}{2} + \frac{\sigma_{\text{xx,e}} - \sigma_{\text{yy,e}}}{2} \cdot \cos 2\phi \\ &+ \tau_{xy,e} \cdot \sin 2\phi.\end{aligned} \tag{3}$$

Here, $\sigma_{\text{xx,e}}$ and $\sigma_{\text{yy,e}}$ represent the normal stresses on the x and y axes of the finite element, respectively, contributing both their average and their difference to the formula. Additionally, the formula includes the shear stress $\tau_{xy,e}$ across the plane. The output from the stress monitoring layer, as depicted in Figure 1, thus consists of the stress histories $\sigma_{\phi,\text{e}}$.

In order to apply the Miner rule, as stated in Equation 1, to the stress histories $\sigma_{\phi,\text{e}}$ within the condition monitoring layer depicted in Figure 1, the rainflow counting method is first performed. This method decomposes complex stress histories into a series of simple, reversed stress cycles, each representing an individual stress response that could potentially lead to material fatigue (Schijve, 2009). The output of the rainflow counting method includes the number of stress cycles n at specific stress amplitudes σ_{a} and mean stress levels σ_{m}. Additionally, the SN-Curve, schematically depicted in Figure 1, is crucial for applying the Miner rule (Schijve, 2009). This curve illustrates the relationship between stress amplitude σ_{a} (with mean stress level $\sigma_{\text{m}} = 0$) and the number of cycles to failure N for a given material. It is essential for implementing the Miner rule, which requires knowledge of the cycles to failure N for specific stress amplitudes. Each point on the SN-Curve represents a specific stress level and its corresponding fatigue life or life expectancy in terms of number of cycles.

Given that simple SN-Curves only address fatigue life under conditions of zero mean stress, mean stress correction is crucial for accurate fatigue life monitoring. The stress cycles n at specific stress amplitudes σ_{a} and mean stress levels σ_{m}, as determined by rainflow counting, are subject to mean stress correction, such as the Goodman mean stress correction method (Schijve, 2009). Once the stress amplitudes are corrected, the Miner rule, shown in Equation 1, can be applied. To assess the LG structural fatigue based on the stress histories $\sigma_{\phi,\text{e}}$ from the stress monitoring layer, the rainflow counting, mean stress correction, and Miner rule must be conducted for the stress histories on every plane at angle ϕ for each finite element e. Consequently, Equation 1 is extended to:

$$D_{\phi,\text{e}} = \sum \frac{n_{\phi,\text{e,i}}}{N_i}. \tag{4}$$

The maximum damage or fatigue D_{LG} experienced by the LG is calculated as follows:

$$D_{\text{LG}} = \max_{\phi,\text{e}} \left(D_{\phi,\text{e}} \right). \tag{5}$$

While calculating the maximum LG fatigue D_{LG} is critical, it is equally important to account for uncertainties in material performance and load observation, as outlined in (Schmidt, 2021). This consideration is implemented using a scatter factor (SF), which, for large civil aircraft, is a minimum of 3, corresponding to material properties with 99 % probability of survival and a 95 % confidence level, as specified by (European Union Aviation Safety Agency, 2020). Therefore,

the safe-life fatigue index $\mathrm{FI_{LG}}$ of the LG is calculated by:

$$\mathrm{FI_{LG}} = \mathrm{SF} \cdot D_{\mathrm{LG}}. \tag{6}$$

This health index serves as the health indicator within the condition monitoring layer of the framework.

The fatigue index, along with data from the usage, loads, stress, and condition monitoring blocks depicted in Figure 1, is stored in a fatigue monitoring database. This database ensures the availability and traceability of all information pertinent to the fatigue monitoring process. For enhanced traceability, it also includes additional information not shown in Figure 1, such as log cards detailing component removals. In conclusion, this database can be integrated into a digital twin or selectively transfer specific data to other systems.

2.2. Fatigue Prediction

The fatigue monitoring framework can be extended by incorporating a prognostics layer, as schematically depicted in Figure 2. Taking the fatigue index $\mathrm{FI_{LG}}$ as input, the RUL calculation

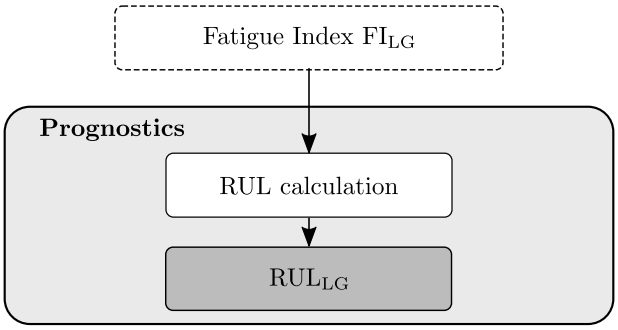


Figure 2. Schematic presentation of LG fatigue prognostics

is straightforward and requires only one main calculation step. Based on the remaining fatigue life estimations by (JIAO et al., 2018), the RUL of the LG is determined by

$$\mathrm{RUL_{LG}} = \frac{1 - \mathrm{FI_{LG}}}{\mathrm{SF} \cdot d} = \frac{1 - \mathrm{SF} \cdot D_{\mathrm{LG}}}{\mathrm{SF} \cdot d}, \tag{7}$$

where d is the predicted mean damage rate in subsequent service. If there is no difference in subsequent service expectable, then $d = 1$. The parameter $\mathrm{RUL_{LG}}$ indicates how much remaining life is left relative to 1, where a value of 1 corresponds to LG failure. To convert this RUL calculation into remaining flight cycles, the equation can be extended by the overall flight cycles n_{FC} experienced by the LG to predict the RUL in terms of remaining flight cycles $\mathrm{RUL_{LG,FC}}$:

$$\mathrm{RUL_{LG,FC}} = \frac{1 - \mathrm{SF} \cdot D_{\mathrm{LG}}}{\mathrm{SF} \cdot d} \cdot \frac{n_{\mathrm{FC}}}{D_{\mathrm{LG}}}. \tag{8}$$

Due to the various sources of uncertainties the precise determination of especially the scatter factor is demanding. The literature provides suggestions (Schmidt, 2021) but a probabilistic estimation of the scatter factor regarding the specific use should be performed when possible.

3. Base model for synthetic data generation

The development of monitoring methods typically requires some sort of data for evaluation. In this case, to assess the model-based LG loads observer, a combination of in-service data recorded by the quick access recorder and dedicated LG loads data is necessary. For this work, method development and evaluation should focus on a narrow-body airliner model with around 100-180 passengers serving as the reference aircraft.

However, due to the unavailability of in-service data and recorded dedicated LG loads, it is essential to generate plausible synthetic in-service and LG loads data. To achieve this, a base model was created using MATLAB/Simulink and the integrated library Simscape Multibody. Simscape Multibody facilitated the implementation of aircraft and LG components within a multibody simulation environment and provided seamless integration with Simulink.

The overall base model consists of the multibody LG model, the airframe, a runway and tyre model as well as an aircraft movement and control subsystem. Figure 3 provides a visualization of the basic model in Simscape Multibody. Further details regarding the structure of the base model are described in the subsequent section.

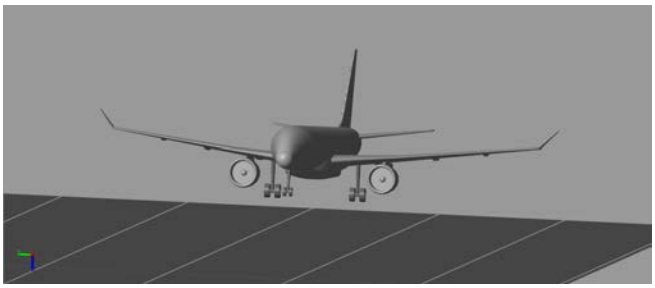


Figure 3. Visualization of the base model: multibody LG model, airframe and runway model

3.1. Multibody landing gear model

The implementation of the multibody LG model was based upon industrial design data, which was made available within the research project OBSERVATOR. Figure 4 illustrates the schematic representation of the implemented bodies and joints for a single main LG in the multibody model. The connection between the aircraft/airframe (AC) and the main fitting (MF) is modeled as a fixed joint (with no degrees of freedom) to represent the LG in an extended and locked state. Given that only load measurements at the LG wheel axle midpoint are of interest, as specified in Section 2.1 due to only one load

application point in the FE assembly model, no additional components connecting the LG and the aircraft, such as side stays, are modeled. However, the impact of these omitted components on LG flexibility is still addressed by integrating their flexibility into the overall LG flexibility matrices which are introduced later in this work.

For simulating translational shock absorber movement, a prismatic joint is installed between the MF and the sliding tube (ST), providing one translational degree of freedom. This design choice simplifies the multibody assembly, obviating the need for additional torque links to prevent ST rotation relative to MF along the rotational axis. Despite this design simplification, the loads calculation at the wheel axle midpoint is not affected. Moreover, to emulate LG flexibility, a single 6-DOF joint is utilized, condensing the LG flexibility into a single flexible point at the wheel axle midpoint. Additionally, revolute joints are employed to constrain the movement of LG wheels W1 and W2 to one rotational degree of freedom each.

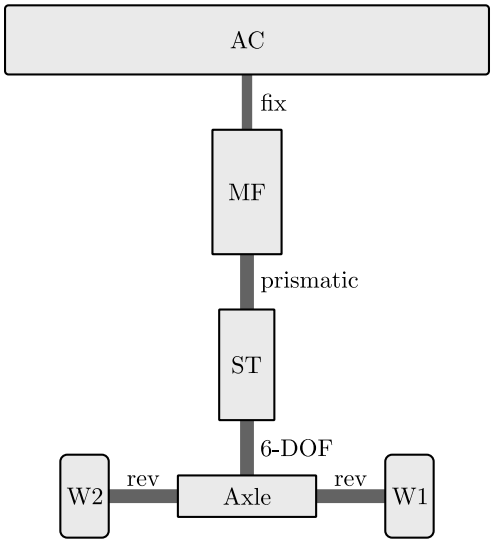


Figure 4. Schematic representation of the bodies and joints of a single main LG

One of the key components of the modeled nose LG configuration is the oleo-pneumatic shock absorber, which primarily provides spring suspension and damping of impact and recoil energy (Schmidt, 2021). To represent the vertical shock absorber dynamics, the shock absorber force, defined by

$$F_{\mathrm{SA}} = F_{\mathrm{spring}}\left(s_{\mathrm{SA}}, T_{\mathrm{amb}}\right) + F_{\mathrm{damp}}\left(s_{\mathrm{SA}}, \mathrm{sgn}(v_{\mathrm{SA}}), v_{\mathrm{SA}}^2\right) + F_{\mathrm{fric}} + F_{\mathrm{limit}}\left(s_{\mathrm{SA}}, v_{\mathrm{SA}}\right) \quad (9)$$

was implemented. Here, F_{spring} represents the force exerted by the gas spring, dependent on the shock absorber travel s_{SA} and ambient temperature T_{amb}. The term F_{damp} is a function of the shock absorber travel s_{SA}, shock absorber velocity $v_{\mathrm{SA}} = \dot{s}_{\mathrm{SA}}$, and the direction of velocity $\mathrm{sgn}(v_{\mathrm{SA}})$, reflecting the oil-induced damping force. Both, the gas spring and the damping force are modelled by the application of lookup tables. Additionally, the shock absorber force accounts for the friction force F_{fric} at the upper and lower bearings of the sliding tube by using simple friction coefficients, along with the translational limiting forces F_{limit} at the upper and lower stops of the shock absorber travel. These upper and lower limiting forces F_{limit} are modelled as simple spring-damper elements, dependent on s_{SA} and v_{SA}.

To implement a flexible LG model, the matrix equation of motion commonly employed in FE analysis was utilized in the LG model:

$$M\ddot{u} + C\dot{u} + Ku = F. \quad (10)$$

Here, F denotes the applied forces and the vector u represents the degrees of freedom of the FE model. M, C, and K denote the mass, damping, and stiffness matrices respectively. Due to computational complexity reduction reasons, only mass, damping and stiffness matrices of the order of 5 were available. With these system matrices reduced by the Guyan model order reduction method (GUYAN, 1965), the LG motion due to flexible structures could be simplified to only one point at the wheel axle midpoint. The computed LG motions were accurately replicated in the multibody model using the depicted 6-DOF joint in Figure 4. However, one translational degree of freedom along the shock absorber axis was disregarded due to the predominant shock absorber travel, leading to the utilization of only 5 degrees of freedom of the 6-DOF joint in Simscape Multibody.

What also had to be taken into account was the change in flexibility with varying shock absorber travel, so that Equation 10 changed to

$$M\left(s_{\mathrm{SA}}\right) \cdot \ddot{u} + C\left(s_{\mathrm{SA}}\right) \cdot \dot{u} + K\left(s_{\mathrm{SA}}\right) \cdot u = F. \quad (11)$$

This implementation issue was addressed by the usage of lookup tables as a function of the shock absorber travel in MATLAB/Simulink. The continuously calculated vector u of Equation 11 could then be input to the 6-DOF joint.

3.2. Tyre model

Tyres represent an essential component of vehicle dynamics such as aircraft LG dynamics. The forces and moments acting on the tyres during ground interaction greatly influence the vehicles dynamics. Thus, when developing multibody LG models, the tyre ground interaction has to be sufficiently represented by tyre models. In contrast to the multibody LG model, data for tyre modelling was not available. This proved to be a challenge, because tyre models in general rely on extensive input parameters. To address this issue, Fiala tyre models were chosen for modelling.The Fiala model is based on a brush-type tyre model and comes with the advantage, that it only requires 10 input parameters which are directly linked to physical properties of the tyre. Due to the fact, that Fiala tyre models for other aircraft types were available, the parameters of those

models could be used for parameter scaling so that plausible assumptions concerning the Fiala tyre models parameters could be made. Nevertheless, the usage of the Fiala model also comes with certain drawbacks as illustrated in (Blundell & Harty, 2004):

- Combined cornering and braking or cornering and accelerating is not considered in the model.
- Aligning moment and lateral force induced by the camber angle are not modelled.
- Varying cornering stiffness at zero slip angle with tyre load is not represented.
- At zero slip angle the offsets in lateral force or aligning moment due to conicity and ply steer are not considered.

Nevertheless, the Fiala tyre model represents a sufficiently good model for the usage of synthetic data generation for landing loads observation model evaluation. The exact mathematical representation of the model, which was used for implementation, is presented in (Blundell & Harty, 2004). The resulting forces and moments, calculated in MABLAB/Simulink, were used as inputs at the contact patches of the individual tyres in the multibody model.

3.3. Runway model

In order to simulate different tyre ground interactions for synthetic data generation, two different runways have been implemented in Simulink and visualized in Simscape Multibody:

- Even runway: A completely even runway with no bumps for optimal landing and taxiing conditions.
- San Francisco Runway 28R: The San Francisco Runway 28R before it was resurfaced was known for high loads on aircraft (European Union Aviation Safety Agency, 2020).

Both runway profiles were constructed using lookup tables in MATLAB/Simulink. The profile of San Francisco Runway 28R was developed based on specifications outlined in (European Union Aviation Safety Agency, 2020). Due to the lack of additional runway data, only these two profiles were employed for synthetic data generation. Within the simulation model, both runway profiles were linked with the tire models of each wheel to simulate tire-ground interaction. Figure 3 provides a visualization of a segment of the even runway.

3.4. Aircraft model and control

The aircraft, or airframe, was modelled as a single rigid body with specific mass and inertia properties. Since no information was available regarding the flight mechanics of similar-sized aircraft, aircraft movement was implemented using forces and moments primarily applied at the aircraft's center of gravity. By incorporating a six-degree-of-freedom joint at the aircraft's center of gravity, the aircraft could be maneuvered along all six degrees of freedom with the multibody LG model mounted on it. To simulate various landing scenarios, multiple controllers were developed. These controllers utilize the forces and moments acting on the aircraft as control variables, along with the aircraft's Euler angles and approach speeds in horizontal, lateral, and vertical directions as reference signals.

3.5. Synthetic data generation for different landing scenarios

The aim of this work, as mentioned in Section 1, is to present a method and evaluate it for observing landing loads in the context of LG fatigue monitoring and RUL prediction. Consequently, the generation of in-service data for various landing scenarios and the recording of dedicated LG loads were required. Simulated landing scenarios included level landings, one-gear landings, side load landings, and rebound landings. To create diverse landing conditions, different parameters were varied. These varied simulation parameters and their value ranges are outlined in Table 1. The variation limits represent plausible assumptions, partly based on knowledge of these parameters from similar aircraft or regulatory documents such as (European Union Aviation Safety Agency, 2020). For the variation of sensor sample rates, only rates up to 50 Hz are theoretically necessary, as higher sample rates are uncommon for aircraft quick access recorder data in today's commercial aviation industry. Nevertheless, additional in-service data sets with a sample rate of 200 Hz were recorded to assess the impact of higher sensor sample rates on monitoring performance.

Table 1. Overview of simulation parameter variations for synthetic data generation

Simulation parameter	**Variation limits**
Roll angle	± 5 deg
Pitch angle	$3 - 9$ deg
Yaw angle	± 5 deg
Aircraft mass	$50,000 - 60,000$ kg
Center of gravity	$22 - 28$ % MAC
Ground speed	$55 - 65$ m/s
Sinking speed	$0.3 - 3$ m/s
Sample rate (in-service data)	$20/50/200$ Hz
Measurement noise (in-service data)	no noise / white noise
Lift force	$0 - 1g$ (variable during touchdown)
Runway profile	even / San Francisco Runway 28R

4. Model-Based Loads Observation of Landing Gear Loads

The following section describes the model-based loads observer of LG loads at the initial landing impact. At first, the developed method is described. In a second step, results of the loads observation method are analysed and the method is evaluated.

4.1. Methodology

Aircraft model-based load observer approaches often rely on a Luenberger observer using specific system sensor data for state estimation, as demonstrated in (Montel & Thielecke, 2018) or (Luderer & Thielecke, 2022). Typically, in-service data recorded at the LG is limited to weight-on-wheel binary signals and rotational wheel speeds. However, employing state estimation within a Luenberger observer with feedback solely based on the the mentioned signals as the only LG signals is not feasible. Therefore, direct estimation of the LG dynamics and loads without state estimation feedback is utilized. A schematic representation of this method using a block diagram and a flow chart is depicted in Figure 5.

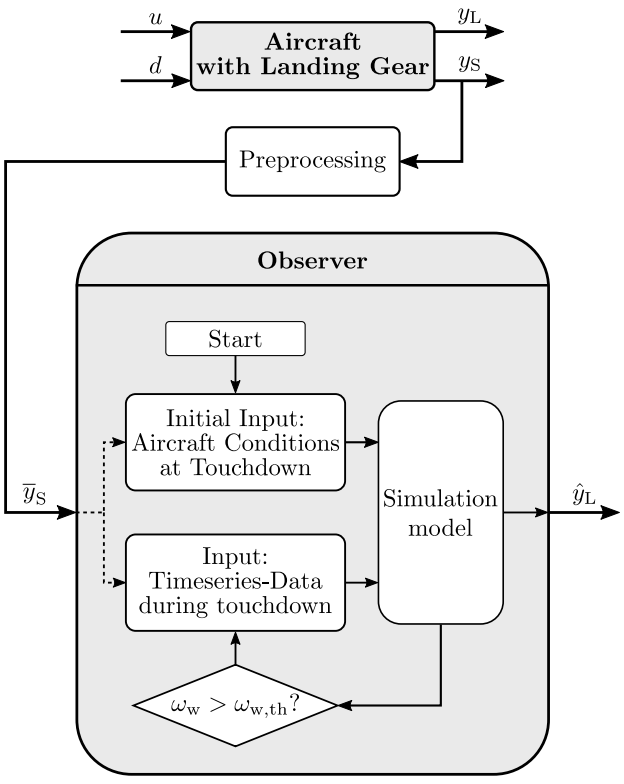


Figure 5. Block diagram of loads observer with data flow of schematic observer logic

During the landing phase and ground operations, the aircraft with the extended and locked LG is controlled with the aircraft inputs u and is simultaneously exposed to various external disturbances d. This results in various loads acting on the aircraft, and particularly in this work, loads y_L acting on the LG. The aircraft and LG dynamics are recorded by various sensors. The recorded in-service data y_S is preprocessed, as described in Section 2.1, so that the preprocessed data $\overline{y}_S$ can be used as loads observer input. The actual loads observer then tries to re-simulate the exact aircraft motion in an offboard simulation from the moment the rotational wheel speed ω_w of at least one main LG wheel exceeds the specified rotational wheel speed threshold $\omega_{w,th}$.

The loads observer begins simulating aircraft movement just above the runway. Initial inputs include the aircraft's roll and pitch angles and approach speeds recorded when the weight-on-wheel signal first changes to 'true' during touchdown. The observer simulates until the simulated rotational wheel speed ω_w of at least one main LG wheel exceeds the specified threshold $\omega_{w,th}$. After this point, the observer uses recorded time series data of longitudinal, vertical, and normal accelerations, roll, pitch, and yaw rates, as well as roll, pitch, and heading angles to reproduce the aircraft movement. The simulated loads are then output by the observer as the signal $\hat{y}_L$, as depicted in Figure 5.

4.2. Analysis

To assess the effectiveness of the proposed observer methodology, it was applied to various synthetic in-service data sets generated by the base model introduced in Section 3, with simulation parameters varied as detailed in Table 1. Initially, baseline simulations were conducted at a sample rate of 200 Hz on an even runway without measurement noise. This setup aimed to exclude potential influences such as uneven runways and low sensor sample rates, allowing for an analysis of the method's performance under 'ideal' conditions. Subsequently, the observer method was tested under more realistic conditions, including measurement noise, uneven runways, and sample rates of 20 Hz and 50 Hz.

A modified version of the base model was used as the simulation model for the model-based loads observer. After baseline and observer simulations, the simulated forces and moments at the main LG wheel axle midpoint during the initial load impact were compared. All observer simulations yielded highly accurate results, with deviations between the simulated baseline and observer loads being less than 2 %. Figure 6 illustrates the forces F_x and F_z at the wheel axle midpoint for a level landing scenario. Here, F_x denotes the force in the longitudinal direction, while F_z indicates the force in the vertical direction of the LG body-fixed coordinate system. The results show that the method performs well for various landing scenarios under conditions of no parameter uncertainties, no measurement noise, an even runway, and high sensor sample rates of 200 Hz or greater.

Subsequently, various parameters of the observer model were individually modified with plausible assumptions to account

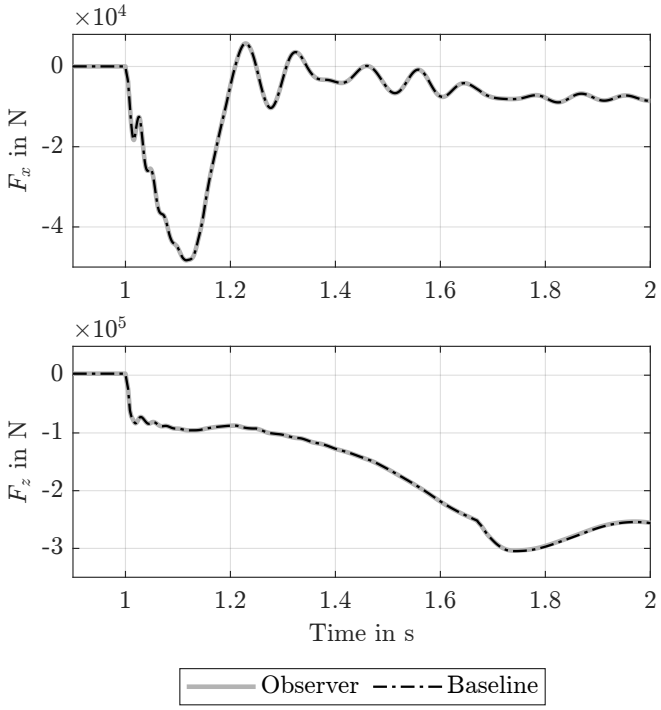


Figure 6. Load observation of longitudinal and vertical forces at wheel axle midpoint of one main LG at touchdown

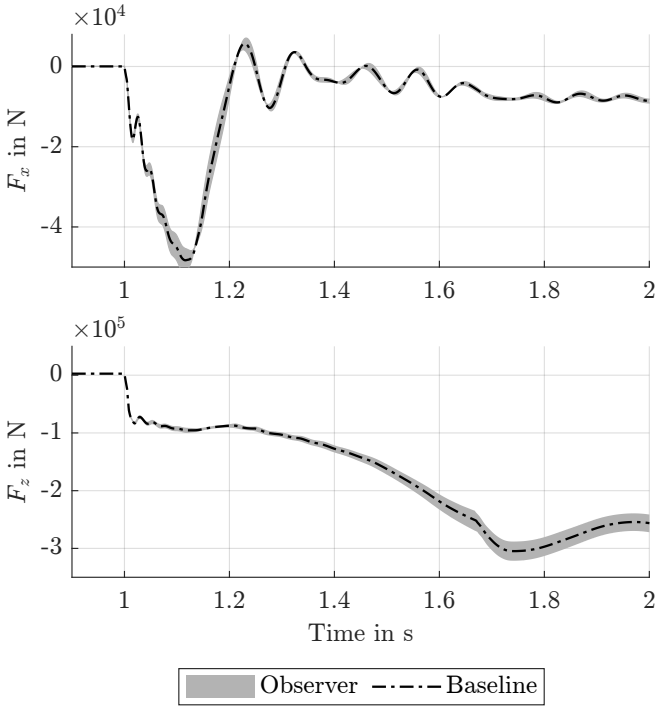


Figure 7. Load observation of longitudinal and vertical forces at wheel axle midpoint of one main LG at touchdown with shock absorber temperature uncertainty of $\pm 10^\circ$ C at 30° C

for observer model uncertainties. The simulations were also conducted with high sample rates of 200 Hz and even runways to avoid biases in the analysis of model uncertainties. The varied parameters include LG flexibility matrices, shock absorber temperature uncertainty, shock absorber spring lookup table, shock absorber damping lookup table, sensor signal offsets, sensor positions, tyre friction coefficients, and tyre stiffness and damping coefficients. Despite these model uncertainties, the deviations between the baseline LG loads and the observer LG loads for the initial landing impact were less than 10 % and were therefore deemed sufficient. For example, Figure 7 illustrates the load estimation bandwidth of the observer for $\pm 10^\circ$ C at 30° C shock absorber temperature uncertainty.

At the time of writing, the impact of the deviations between baseline and observer loads in the fatigue monitoring framework introduced in Section 2.1 is not fully known. Therefore, it is not yet possible to make exact statements about the quality of the observer results. Nonetheless, the initial findings indicate a potential for precise load monitoring despite model uncertainties.

Furthermore, the influence of sensor sample rates on the observer method has been examined. For example, the load observations of the longitudinal and vertical forces at the wheel axle midpoint are depicted for different sensor sample rates in Figure 8. While the observer performs well for load observation with sensor sample rates of 200 Hz, the quality of load estimation decreases with decreasing sample rate.

Figure 8 also reveals that both the observer with 20 Hz sample rate and the one with 50 Hz sample rate start to diverge from the baseline in F_z at approximately the same time. This occurrence can be attributed to a significant increase in the vertical deceleration of the aircraft about the same time, leading to imprecise recordings of vertical accelerations during observer simulations. Nevertheless, sample rates of 50 Hz, which are common in modern aircraft, still hold considerable potential for effective observation of LG loads during the initial landing impact at the main LGs for use in LG fatigue monitoring.

Another significant factor expected to influence load observation was landing on uneven runways. Baseline simulation results are depicted in Figure 9. These show the exemplary longitudinal and vertical forces at the wheel axle midpoint for a level landing scenario on the San Francisco Runway 28R profile (before resurfacing). This runway was known for inducing high loads due to its uneven nature. The figure also presents the corresponding observer loads simulated for an even runway, as the observer lacked information about the actual runway profile. An observer sample rate of 200 Hz was employed to mitigate potential inaccuracies from inadequate sample rates, thus excluding certain erroneous load estimations. At the beginning of the landing impact, when aircraft movement predominates and no critical runway bumps affect the LG, the observer estimates the LG loads quite accurately, albeit with higher frequency oscillations. However, as simulation time progresses, the loads begin to deviate significantly

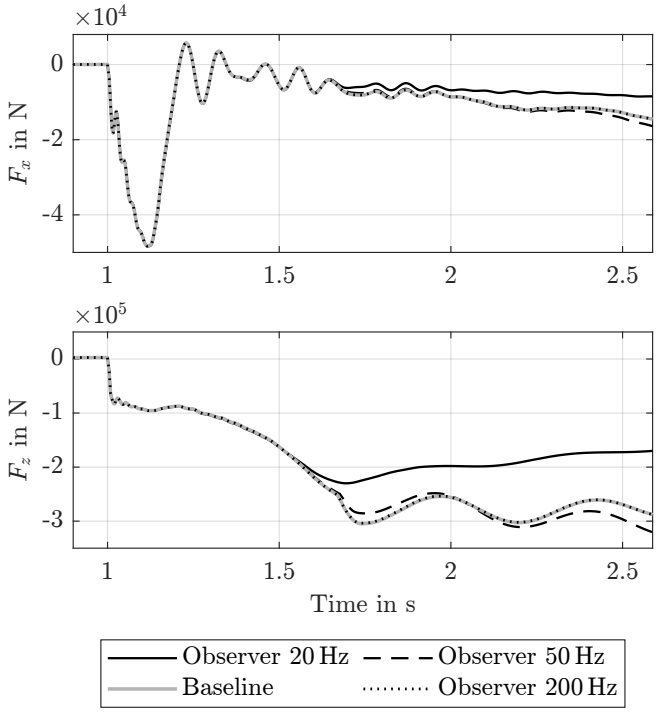


Figure 8. Loads observation of longitudinal and vertical forces at wheel axle midpoint of one main LG at touchdown for different virtual in-service data sample rates

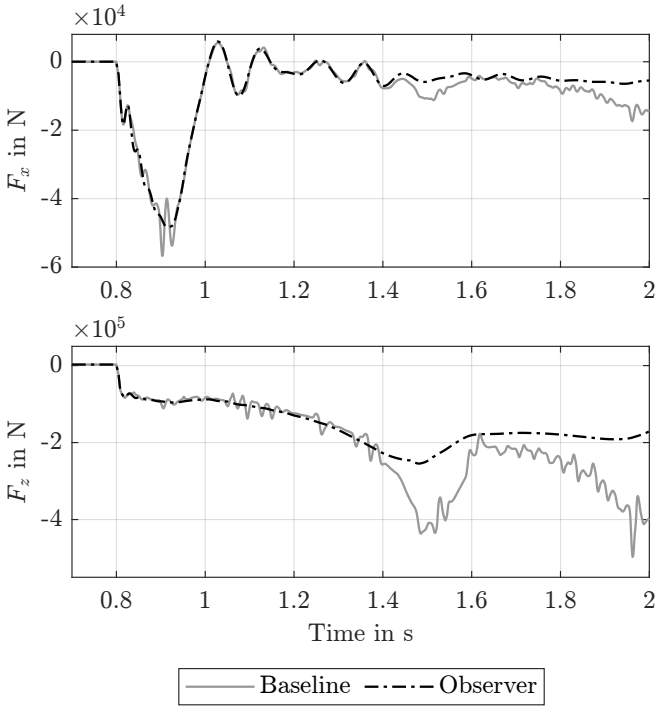


Figure 9. Loads observation of longitudinal and vertical forces at wheel axle midpoint of one main LG at touchdown: baseline touchdown on San Francisco Runway 28R profile, observer touchdown on even runway

due to runway bumps and variations in runway height.

The estimation results of the observer could be significantly improved by observer simulations with a known uneven runway profile and a known runway position of the aircraft during touchdown. However, if the precise landing position and especially the runway surface profile are unknown, which is usually the case nowadays, highly uneven runways can lead to significant variations in runway excitation. Despite maintaining the same vertical aircraft position in the observer as in the baseline, deviations in loads can be substantial due to these discrepancies.

5. Conclusion

This paper introduces a model-based LG loads observer method that operates exclusively on in-service data, thereby eliminating the need for additional sensors. The method is specifically evaluated with an emphasis on the first landing impact of the main LGs. It forms a key component of a comprehensive LG fatigue monitoring framework and the subsequent calculation of RUL for the LG. This paper also fundamentally outlines the foundational steps and further key components for LG fatigue monitoring and prediction, based on the 'safe life' design methodology commonly used for structural LG certification.

The application of the loads observer method on virtual in-service data with dedicated LG loads shows significant potential despite challenges, such as the unsuitability of LG feedback signals for state feedback observers and the heavy influence of sample rates on precision. For accurate load estimation, particularly for initial landing impacts, sample rates of at least 50 Hz are necessary. However, deviations between recorded and actual aircraft accelerations can lead to unacceptable estimation errors over time, suggesting that higher sample rates might be needed for longer monitoring durations.

A major challenge is the unknown runway profile, notably on uneven runways like the pre-resurfaced San Francisco Runway 28R, where load estimation accuracy drops significantly. The position inaccuracies in the observer model, due to integrating recorded aircraft accelerations, further distort load estimations on inclining runways. Precise runway profiles and exact touchdown coordinates are crucial for improving estimation accuracy.

While this paper demonstrates a basic proof of concept by applying the developed method to virtual in-service data, further analysis and development are required to address the challenges associated with load estimation. For instance, the exact effects of load estimation errors on fatigue and RUL determination need to be investigated. Additionally, knowledge about runway profiles and the exact touchdown position must be incorporated. Furthermore, combined effects of model uncertainties on load observation should be explored.

ACKNOWLEDGMENT

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action within the project OBSERVATOR (contract code: 20D1903C) in the national LuFo program. Their support is greatly appreciated.

REFERENCES

Blundell, M., & Harty, D. (2004). *Multibody systems approach to vehicle dynamics*. Amsterdam: Elsevier Butterworth-Heinemann.

Boller, C., & Buderath, M. (2007). Fatigue in aerostructures — where structural health monitoring can contribute to a complex subject. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1851), 561–587. doi: 10.1098/rsta.2006.1924

Boller, C., & Staszewski, W. J. (2004). Aicraft structural health and usage monitoring. In W. J. Staszewski, C. Boller, & G. R. Tomlinson (Eds.), *Health monitoring of aerospace structures.* Chichester: John Wiley.

Buderath, M. (2009). Fatigue monitoring in military fixed-wing aircraft. In C. Boller, F.-K. Chang, & Y. Fujino (Eds.), *Encyclopedia of structural health monitoring.* Chichester: Wiley.

Buderath, M., & Neumair, M. (2007). Operational risk assessment for unmanned aircraft vehicles by using structural health and event management. *UAV Design Processes / Design Criteria for Structures*, 2.1-1 – 2.1-10.

Chabod, A. (2022). Digital twin for fatigue analysis. *Procedia Structural Integrity*, *38*, 382–392. doi: 10.1016/j.prostr.2022.03.039

Dziendzikowski, M., Kurnyta, A., Reymer, P., Kurdelski, M., Klysz, S., Leski, A., & Dragan, K. (2021). Application of operational load monitoring system for fatigue estimation of main landing gear attachment frame of an aircraft. *Materials (Basel, Switzerland)*, *14*(21). doi: 10.3390/ma14216564

El Mir, H., & Perinpanayagam, S. (2021). Certification approach for physics informed machine learning and its application in landing gear life assessment. In *2021 ieee/aiaa 40th digital avionics systems conference (dasc)* (pp. 1–6). IEEE. doi: 10.1109/DASC52595.2021.9594374

El Mir, H., & Perinpanayagam, S. (2022). Certification of machine learning algorithms for safe-life assessment of landing gear. *Frontiers in Astronomy and Space Sciences*, *9*. doi: 10.3389/fspas.2022.896877

European Union Aviation Safety Agency. (2020). Certification specifications and acceptable means of compliance for large aeroplanes cs-25: Amendment 26.

GUYAN, R. J. (1965). Reduction of stiffness and mass matrices. *AIAA Journal*, *3*(2), 380. doi: 10.2514/3.2874

Holmes, G., Sartor, P., Reed, S., Southern, P., Worden, K., & Cross, E. (2016). Prediction of landing gear loads using machine learning techniques. *Structural Health Monitoring*, *15*(5), 568–582. doi: 10.1177/1475921716651809

Jeong, S. H., Lee, K. B., Ham, J. H., Kim, J. H., & Cho, J. Y. (2020). Estimation of maximum strains and loads in aircraft landing using artificial neural network. *International Journal of Aeronautical and Space Sciences*, *21*(1), 117–132. doi: 10.1007/s42405-019-00204-2

JIAO, R., HE, X., & LI, Y. (2018). Individual aircraft life monitoring: An engineering approach for fatigue damage evaluation. *Chinese Journal of Aeronautics*, *31*(4), 727–739. doi: 10.1016/j.cja.2018.02.002

Lee, Y.-L., & Barkey, M. E. (2012). Stress-based multiaxial fatigue analysis. In Y.-L. Lee, M. E. Barkey, & H.-T. Kang (Eds.), *Metal fatigue analysis handbook.* Waltham, Mass: Butterworth-Heinemann.

Luderer, O., & Thielecke, F. (2022). Validation of a hybrid loads observer for a subscale test aircraft with distributed electric propulsion. *33rd Congress of the International Council of the Aeronautical Sciences ICAS*.

Montel, M., & Thielecke, F. (2018). Validation of a hybrid observer method for flight loads estimation. *31st Congress of the International Council of the Aeronautical Sciences, ICAS 2018*.

SAE International. (2020). *Landing gear fatigue spectrum development for part 25 aircraft: Air 5914.* 400 Commonwealth Drive, Warrendale, PA, United States: Author. doi: 10.4271/AIR5914

Schijve, J. (2009). *Fatigue of structures and materials* (Second edition ed.). Dordrecht: Springer.

Schmidt, R. K. (2021). *The design of aircraft landing gear*. SAE International. doi: 10.4271/9780768099430

Schmidt, R. K., & Sartor, P. (2009). Landing gear. In C. Boller, F.-K. Chang, & Y. Fujino (Eds.), *Encyclopedia of structural health monitoring.* Chichester: Wiley.

Stolz, C., & Neumair, M. (2008). Structural integrity management system for enhanced aircraft availability. *Future Airframe Structural Lifing: Methods, Applications and Management*, 24-1 - 24-12.

Tinga, T. (2010). Application of physical failure models to enable usage and load based maintenance. *Reliability Engineering & System Safety*, *95*(10), 1061–1075. doi: 10.1016/j.ress.2010.04.015

Process Quality Monitoring Through a LSTM Network Derived from a Rule-Based Approach

Andreas Bernroitner[1] and Roland Eckerstorfer[2]

[1,2] *Plasser und Theurer, Linz, 4020, Austria*
andreas.bernroitner@plassertheurer.com
roland.eckerstorfer@plassertheurer.com

ABSTRACT

The railway infrastructure condition is a crucial factor for the safe and efficient operation of trains. Regular maintenance is inevitable as the track geometry degrades over time due to traffic and environmental effects. To restore the ideal position and provide sufficient durability of ballasted track so called tamping machines are used. These machines lift the track, correct the longitudinal level and the alignment of the track panel and tamp the ballast. During the tamping process the tamping tines penetrate the ballast bed, fill voids and compact the ballast underneath the sleepers by a squeezing movement with superimposed vibration. A detailed description of the tamping cycle can be found on section 2. Monitoring and evaluating this tamping process is essential for maintaining process quality. This can be achieved through a variety of sensors, such as incremental encoders, angle encoders, temperature, pressure, and acceleration sensors, coupled with a measurement unit (DAQ and edge device) to collect, locally store and transmit the data to a cloud. This paper explores the development of a rule-based algorithm for assessing the quality of the tamping process execution in reference to its nominal chronological sequence. The focus is on identifying tamping occurrences and classifying them into acceptable (OK) or non-acceptable (NOK) categories. This involves selecting relevant measurement parameters and processing them, considering the inherent imprecision in real-world processes. Empirical thresholds are established to differentiate between good and bad outcomes. The classification approach has to be sufficiently generic in order to cover a high variety of customized tamping machine types. As each machine is individually designed, the process of generalization is challenging and complex. The paper demonstrates the accuracy and universal applicability of the developed rule set across different tamping machines. The model's effectiveness is validated using the Hold-Out-Test-Set method. Furthermore, the rule-set-achieved outcomes are compared with results gained from an LSTM network. Both the rule-based approach and the neural network demonstrate precision, but the latter requires significantly more effort.

1. TRACK MAINTENANCE

For a safe and efficient operation of trains a proper track infrastructure is indispensable. Especially on high-speed rail links the quality of the track and its surroundings is crucial. Therefore not only the construction but also the maintenance of the track in order to prevent degradation due to traffic and environmental effects are important. This involves ensuring a clean and dry embedding, sufficient proper ballast underneath the rails, impeccable condition of the sleepers involved and restoring the vertical as well as the horizontal position of the rails. A very detailed description of several track maintenance methods can be found in (Hansmann, 2021).

In Figure 1 an acceptable condition of a track is depicted. Here a sufficient amount of appropriately sized, clean ballast is in place. The positioning of the rails in vertical and horizontal direction is within applicable limits. The durability of the track geometry is ensured through appropriate compaction of the ballast.

Figure 1. Acceptable condition

For the latter tasks so-called tamping machines are used. The process which results in appropriately compacted ballast is referred to as tamping. In Figure 2 an example for such a tamping machine can be seen.

Figure 2. Tamping machine

A crucial part of such working machines is the tamping unit which is visualized in Figure 3. The lower grey colored components are called the tamping tines. They constitute the only components which are in direct contact with the ballast.

Figure 3. Tamping unit

Ultimately, the focus of this paper is the automatized identification of tamping cycles. Subsequently also classifying tamping cycles into acceptable and non-acceptable cycles, hereinafter denoted as OK and NOK respectively, will be done.

In Figure 4 a track with an unacceptable positional deviation can be seen. The ballast condition regarding size, homogeneity and cleanness does not fulfil the minimum criteria either.

Figure 4. Unacceptable condition of the track

In Figure 5 it is obvious that the ballast is in an unacceptable condition. Neither the ballast size nor the cleanness meet the desired conditions. (Soleimanmeigouni I, Ahmadi A, Kumar U., 2018) provide a summary, discussion and classification of existing track geometry measures and track geometry degradation models. Machine learning approaches for diagnosis and prognosis of rail defects are reviewed by (Chenariyan Nakhaee, Hiemstra, Stoelinga, & van Noort, 2019).

Figure 5. Unacceptable condition of the ballast

2. TAMPING PROCESS

The main goal of tamping is to correct track faults in longitudinal level and alignment in order to guarantee the operating

reliability and ride comfort of the trains. The explanations given in this section are based on (Offenbacher, Koczwara, Landgraf, & Marschnig, 2023) and (Fellinger, 2017). Furthermore, ballast faults like voids beneath the sleepers should be corrected so that the load onto the sleepers is equally distributed and deployed to the underfloor. This also increases track quality before irreversible damages can occur.

A complete tamping cycle can be decomposed into the following sub-processes:

1. Positioning
2. Lifting and Lining
3. Penetrating
4. Filling
5. Compacting
6. Lifting

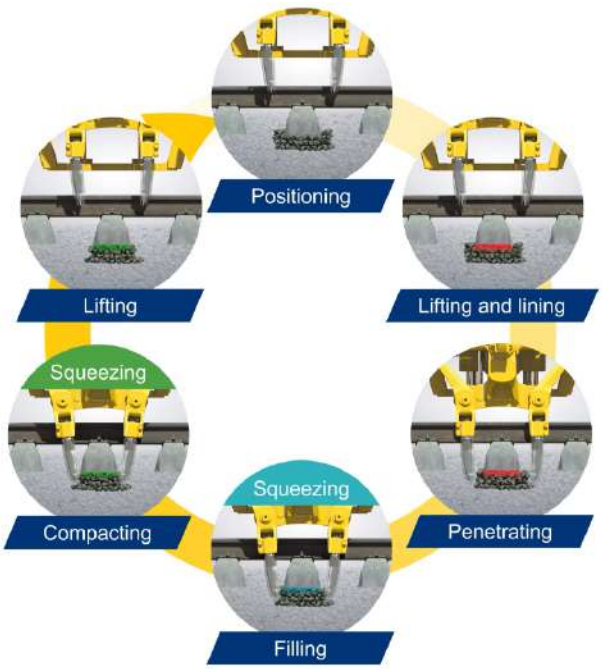


Figure 6. Tamping stages

In the Positioning phase it has to be ensured that the tamping unit is positioned exactly above a sleeper and there is no relative velocity between the unit and the track. Subsequently, in the Lifting and Lining phase the rails are correctly positioned by a separate working unit. Here the rails are lifted and brought into the desired longitudinal and lateral position. Then Penetrating is done and the whole tamping unit is lowered until the tamping tines sink into the surface and the lower position is reached. This is followed by the squeezing movement of the tines which basically comprises two sub processes, Filling (the void caused by the previous Lifting and Lining with ballast) and Compacting (the ballast under the sleeper). Finally, rails are released and the tamping unit is retracted again. This process is known as Lifting. During all of the stages the vibration has to be active. Thus the tamping tines are oscillating with 35 Hz for a smoother penetration and squeezing movement inside the ballast bed (Fischer, 1983).

3. Data Generation/Measurement System

A variety of sensors, such as incremental encoder, angle encoders, temperature, pressure, and acceleration sensors are connected with the control system. An Industrial Internet of Things (IIoT) edge device is fully integrated with the machine control system via the machine network. The device collects and records the data which is transferred to an online platform by means of a mobile broadband connection.

4. Tamping Abstraction

Unfortunately, there are not one-to-one relationships between the recorded measurement signals and the sub-processes as described in Section 2. Additionally, there are further conditions to be fulfilled to assess the quality of the tamping-process, e.g., squeezing (consisting of filling and compacting) shall only be performed when the tamping unit already rests in the down position and not during penetration. On the other hand, it is not relevant to distinguish filling from compacting, but only the process of squeezing and related key parameters as squeezing times are of interest. The sub-process “Positioning” can only be identified by means of the vehicle’s speed, in detail, whether the machine is at standstill or not, but it cannot be checked if it is positioned properly. There are separate assisting tools which deal with proper positioning. For example, there is a camera and image recognition system that makes suggestions to the operator for adjusting the tamping units properly, especially in turnouts. The operator only needs to confirm the suggestions (Plasser und Theurer, 2017). Concluding, the sub-processes as depicted in figure 6 need be represented by sequences based on and created by real signal data. Therefore, the signals are transformed into segments of Boolean representations by means of applying mathematical operations and threshold values, if required. The correct sequence of serial and parallel segments determines the quality or correctness of the tamping process. The proper sequence of segments for an acceptable tamping process is depicted in figure 7.

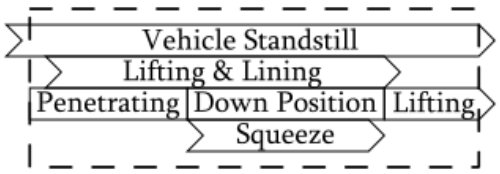


Figure 7. Tamping process based on measurement data

However, real data show deviations from theory, like the overlaps of segments or short unidentified periods between segments (i.e., pauses) which should be in series. Furthermore, differing start or end times of segments which should be synchronous may occur. The inaccuracies are caused by different sampling rates of the individual signals or temporal shifts induced by mathematical operations. Thus, there are parts in the sequences which do not follow the strict theoretical rules but can be considered as valid to a certain extent. The imprecision necessitates the definition of further rules to qualify a tamping process.

5. LONG-SHORT-TERM-MEMORY

The Long Short-Term Memory (LSTM) network was invented by Sepp Hochreiter and Jürgen Schmidhuber in 1997 (Hochreiter & Schmidhuber, 1997). Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) architecture used in the field of deep learning. LSTMs are designed to avoid the long-term dependency problem typical of standard RNNs, enabling them to remember information for long periods. This makes LSTMs particularly useful for tasks involving sequential data, such as time series analysis, natural language processing (NLP), speech recognition, and more. The key to LSTM's ability to retain long-term memory is its cell state, along with its various gates that control the flow of information. An LSTM unit typically comprises the following components:

- Forget Gate f_t
- Input Gate i_t
- Cell State c_t
- Output Gate o_t

The Forget Gate decides what information should be thrown away or kept. It looks at the current input and the previous hidden state and outputs numbers between 0 and 1 for each number in the cell state (C_{t-1}). A value close to 1 means to keep the information, while close to 0 means to forget it. The Input Gate decides what new information will be stored in the cell state. It involves two parts: one Sigmoid layer that decides which values to update, and a Tanh layer that creates a vector of new candidate values that could be added to the state. The cell state is the key innovation of LSTMs. It runs straight down the entire chain, with only minor linear interactions. It's very easy for information to just flow along it unchanged. The cell state is modified by the forget gate and the input gate. The Output Gate determines the next hidden state, which contains information on previous inputs. The hidden state can be used to make predictions. The output gate looks at the current input, the previous hidden state, and the current cell state, and decides what the output should be. These components work together to allow the LSTM to decide when to allow data to enter, when to forget data because it's no longer useful, and when to let it impact the output at the current timestep. This selective memory capability helps LSTMs to perform exceptionally well on tasks where the context or the sequence of data points is important.

A Bidirectional Long Short-Term Memory (Bi-LSTM) network is an extension of the traditional Long Short-Term Memory (LSTM) network. It enhances the original LSTM by providing two layers that process the input sequence in both forward and backward directions. By processing sequences in both directions, Bi-LSTMs can capture context from both the past and the future relative to a specific point in the sequence. The key idea behind a Bi-LSTM is that at any point in time, the network has access to information from both the beginning and the end of the sequence, making it especially powerful for tasks where context from both directions is crucial for understanding or predicting the elements of the sequence. Mathematically, a Bi-LSTM combines the outputs from two separate LSTM layers — one processing the input sequence from start to end and the other processing it from end to start. The outputs of these two LSTMs can be merged in various ways (e.g., concatenation, summation, or averaging) to form a single output that provides a comprehensive context-aware representation of each point in the sequence. Bi-LSTMs are widely used in various sequence modeling tasks, such as natural language processing for named entity recognition, sentiment analysis, and machine translation, as well as in bioinformatics and speech recognition, where understanding the context from both directions can significantly enhance model performance.

In Figure 8 a typical (vanilla-)LSTM is depicted. In the graph the σ stands for the Sigmoid activation and the $\tanh$ for Hyperbolic Tangent activation function. g_t represents the input activation and the x an element wise multiplication.

(De Simone et al., 2023) describe the application of a LSTM model for the failure prediction of rolling stock equipment, in detail of the traction converter cooling system, but also give a rough overview on other LSTM-based prediction algorithms in the railway industry.

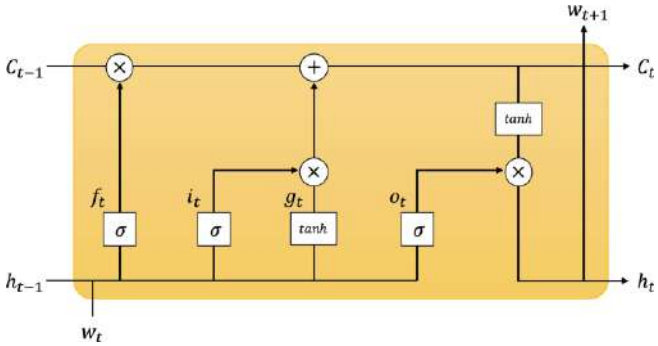


Figure 8. LSTM workflow
(Park & Kim, 2020)

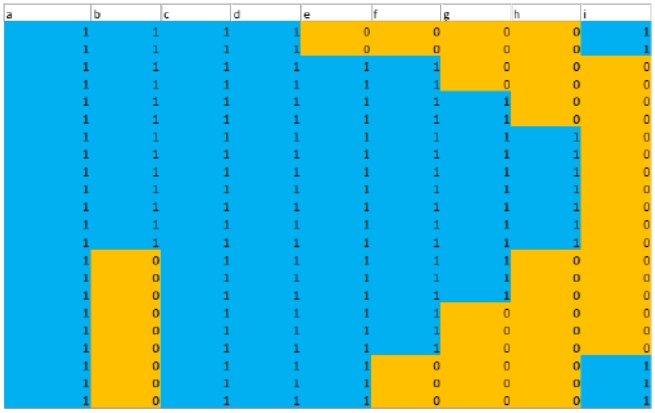


Figure 9. Tamping identification

6. Rule Based Detection

6.1. OK Tamping cycle detection

In order to identify and evaluate tamping cycles based on the time series of the measurement channels, boolean signals are created and assembled, as depicted in Figure 9. A sampling rate of 10 Hz was chosen in order to ensure an appropriate resolution of the signals. Each row of the visualization in Figure 9 is a time increment. Column **a** stands for "engine running", which means that the engine of the tamping machine has to be switched on and a minimum rotational speed has to be exceeded. The second column **b** represents the tamping cycle initialization which is done by the operator by means of a foot-operated pedal. In the **c** column it is listed whether the superimposed vibration of the tamping tines is activated or not. **d** indicates the proper lifting and lining of the rails. In **e** one can see if the tamping unit's relative velocity falls below a very low threshold value with respect to the rails. This means that "the tamping unit stands still" or it is in a very slow movement at least. The penetrating phase is described in column **f** via checking the downward movement of the tamping units, in detail, it is true if it moves and false if not. Column **g** shows if the tines are in the desired lower position. This is again realized by applying a threshold value to the tamping unit's positional encoder. In column **h** the squeezing movement is depicted. It is true if the tamping tines are moved towards each other to fill and compact the ballast under the sleeper and false else. In the last column **i** the retraction, the lifting of the unit, is depicted. For the consideration of measurement and transmission errors small deviations are tolerated. This means that also segments which are disconnected by only one or two time increments are regarded as one full coherent segment.

The following criteria are established for the tamping cycle identification and classification:

- duration of each individual segment
- simultaneity of segments
- duration of sections with overlapping segments which should not be simultaneous
- serial sequence of segments or detachment of consecutive signals
- duration between consecutive segments

The definition of permissible durations, serial sequences, concurrences etc. requires both profound domain knowledge about the tamping cycle and empirical insights based on real data. For example, the ideal minimum squeezing time, i.e., the duration from start of the filling phase until the end of compaction phase, is defined as about 1.2 seconds by a manufacturer of tamping machines. However, there can be national regulations which specify a deviating squeezing duration. Another example can be the temporal succession of the lifting and penetration phases. Ideally, these two sequences are strictly in series. However, the downward movement of the tamping unit can already start when the lifting of the rail is still in progress provided that a void has formed as soon as the tamping tines enter the ballast. Concluding, a certain duration of parallelism is permissible in this case. Furthermore, it is also acceptable that there is a short pause between the segments. Thus, the definition of such thresholds and tolerances requires experience and sensitivity from the engineers and data analysts. Usually the threshold values are determined empirically or are defined be national regulations depending on where the machine is operated.

In order to get an intuitive feeling about the identification process several consecutive tamping cycles are depicted in Figure 10, where blue sections represent boolean true and orange, boolean false. The columns are identical to those in Figure 9.

Figure 10. Tamping identification of multiple cycles

6.2. NOK Tamping cycle detection

In order to obtain the desired process quality all sub-processes have to have the appropriate duration and also the sequence of the consecutive sub-processes has to satisfy the correct order, which means that the subsequent signal has to follow the previous one within a certain amount of time. Therefore the following error scenarios can occur:

- vibration off
- incomplete vibration
- relative velocity
- no stand still
- incomplete stand still
- no leveling
- incomplete leveling
- no penetration
- no down position
- incomplete down position
- no squeezing movement
- penetration before lifting the rails
- squeezing before down position
- lifting the tamping unit before squeezing

The time-series signals of an example of detected NOK tamping cycles are depicted in Figure 11 - a better quality of the plot can be found in the Appendix 9, too - where the upper graph shows the lowering position of a tamping unit. Negative values indicate positions of the tamping tines above the rail, zero is approximately the level of the rail's surface and positions greater than 120 mm can be considered as the tines entering the ballast. The lower graph illustrates the machine's velocity in m/h. Based on the developed cycle identification an impermissible overlap of the two sections "vehicle stands still" and "tamping tines are in the ballast" could be detected. The overlap is highlighted in red colour in Figure 11. These overlaps indicate that the tines are already located in the ballast even though the machine is still moving can cause significant wear on or even severe damage of the tamping unit.

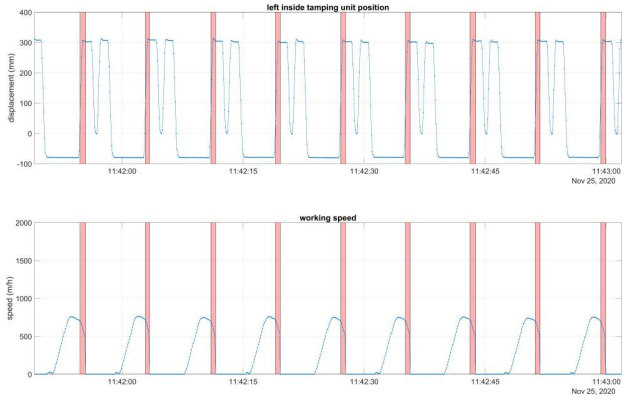


Figure 11. Identified NOK tamping cycle: The tamping unit is already in the ballast even though the machine is still moving (see also Appendix)

7. LSTM Detection

7.1. Architecture

After several trials regarding the structure of the network, the architecture depicted in 9 was chosen.

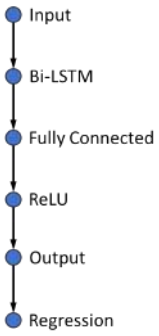


Figure 12. LSTM architecture

The network consists of:

- Input Layer
- Bi-LSTM Layer
- Fully Connected Layer
- ReLU Activation
- Output Layer
- Regression Output

The input layer is the bottom-most layer, where the input, the previously generated boolean signals, is fed to the network.

Subsequently data is passed to the Bi-LSTM layer which consists of 25 hidden units. Bidirectional LSTMs can be useful when the context of the input is needed from both the past and the future of a specific time step. This turned out to be the case in the cycle identification. Following the Bi-LSTM layer, there is a fully connected layer which takes the sequential output from the Bi-LSTM and transforms it into a fixed-size vector. This layer has 10 units, and it is likely responsible for integrating the features learned by the Bi-LSTM layer. After that a non-linearity in form of a ReLU activation function is applied. Therewith the model is allowed to account for non-linear relationships between the features. The next layer is another fully connected output layer with a single unit. This is because the network is designed to output a single continuous regression value. The final layer is a regression output layer with the mean squared error as loss function.

7.2. Training

The analysis workflow was implemented in Matlab and it turned out that training for only 10 epochs with 225 iterations each is sufficient. For the training the timeseries were split into windows of 10 seconds each and a step size of 5 seconds was chosen. Therefore an overlap of 50% occurred intentionally. The training was done with a learning rate of 0.001 on a single GPU. and the training in total only took a couple of minutes. The metric used in training was RMSE (root mean square error).

7.3. Results

In Figure 13 the cycle detection can be seen. This graph can also be found in the appendix. In this visualization the gray rectangles represent the tamping cycles and also their duration. A nearly perfect fit can be found here. There is no visible deviation between the LSTM and the rule based results.

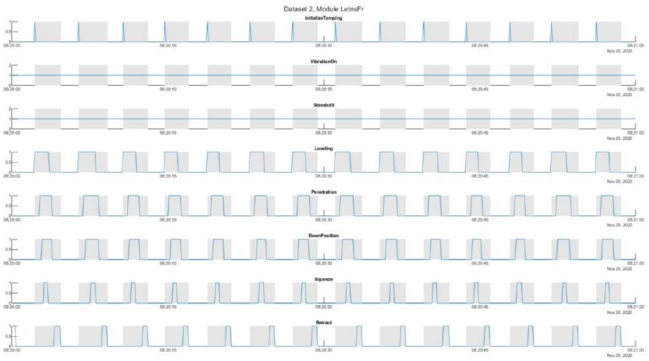


Figure 13. LSTM tamping cycle detection (see also Appendix)

Using the hold-out test set method, an accuracy of 0.98 was achieved.

8. Conclusions

The comparison of the two tested methods for tamping cycle identification, i.e. the rule-based vs. the LSTM approach, it can be concluded that:

1. The accuracy of the rule-based method is approx. 100%, whereas that of the LSTM model is approx. 98% taking only OK detections into consideration. Obviously, the rule-based approach, which basically consists of a set of subsequent if-queries, delivers better results due to the fact that the rules exactly represent the definition of a correct tamping cycle. But the exact representation requires profound domain knowledge of and experience on the tamping procedure and the data acquisition process. When lacking this knowledge and experience the neural network, which defines its own rules by adjusting its learnable parameters, the weights and biases, by evaluating the time series over and over, turns out to be a suitable alternative to still get very accurate results.

2. The implementation effort for the LSTM model is much higher as well as the required hardware and processing resources for training and evaluating the network.

3. The pre-processing of the data and the generation of the boolean sub-processes is the same for both methods.

4. The identification of the NOK tamping cycles is more difficult for the LSTM approach due to the lack of sufficient amount of NOK cycles in real world training data because operating errors rarely occur. A possible solution would be to artificially generate error cases in order to allow the model learn incomplete sequences.

9. Further steps

The LSTM approach as described is capable of identifying OK-cycles. However the NOK-cycles are of higher interest with regards of wear and resulting maintenance. However these cases do not occur sufficiently frequent in real world data. Therefore artificial samples could be generated and be fed to the training set. Another approach could be weighing the very rarely occurring failure cycles higher than the frequently occurring OK cycles in order to balance the training set. Furthermore it should be checked if the found algorithm is generic enough to also fit to other machines and surroundings. Thus it shall be enrolled to different machines operating in different regions of the world in order to compare results and performance subsequently. On the other side also other algorithms shall be implemented and compared. Therefore the time series should again be split into small segments e.g. 0.1s and each of these segments should be classified by different machine learning algorithms according to the features within the respective segment.

REFERENCES

Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M., & van Noort, M. (2019). The recent applications of machine learning in rail track maintenance: A survey. In S. Collart-Dutilleul, T. Lecomte, & A. Romanovsky (Eds.), *Reliability, safety, and security of railway systems. modelling, analysis, verification, and certification* (pp. 91–105). Cham: Springer International Publishing.

De Simone, L., Caputo, E., Cinque, M., Galli, A., Moscato, V., Russo, S., … Giannini, G. (2023). Lstm-based failure prediction for railway rolling stock equipment. *Expert Systems with Applications*, *222*, 119767.

Fellinger, M. (2017). Validierung der instandsetzungsmengen der standardelemente gleis der öbb. *FSV aktuell Schiene*.

Fischer, J. (1983). *Einfluss von frequenz und amplitude auf die stabilisierung von oberbauschotter* (PhD thesis). Technische Universität Graz.

Hansmann, S., Nemetz. (2021). Keeping track of track geometry.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*.

Offenbacher, S., Koczwara, C., Landgraf, M., & Marschnig, S. (2023). A methodology linking tamping processes and railway track behaviour. *Applied Sciences*.

Park, S., & Kim, Y. (2020). A method for sharing cell state for lstm-based language model. *Computer and Information Science*, 81–94.

Plasser und Theurer. (2017). The track maintenance machine 2020. *Plasser und Theurer today*, *46*(132), 24–26.

Soleimanmeigouni I, Ahmadi A, Kumar U. (2018). Track geometry degradation and maintenance modelling: A review. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit.*, *232(1)*, 73-102.

BIOGRAPHIES

Andreas Bernroitner obtained his Master of Science (M.Sc) degree in Mechanical Engineering from the University of Applied Science in 2014, following the completion of his Bachelor of Science (B.Sc) at the same institution. He is presently pursuing a second M.Sc degree in Artificial Intelligence at Johannes Kepler University, Linz. With professional experience as a failure analysis engineer, Bernroitner has been serving as a data scientist within the *Data Science and Analytics* team at *Plasser und Theurer*, Linz, since 2022.

Roland Eckerstorfer graduated in Medical Device Technology on the University of Applied Sciences Upper Austria in Linz in 2009. After several years as measurement engineer in research and development of special purpose machines performing measurements and subsequently the data evaluation and analysis in power plants worldwide and on test rigs, he now is a member of the *Data Science and Analytics* team at *Plasser und Theurer* in Linz since 2021.

APPENDIX

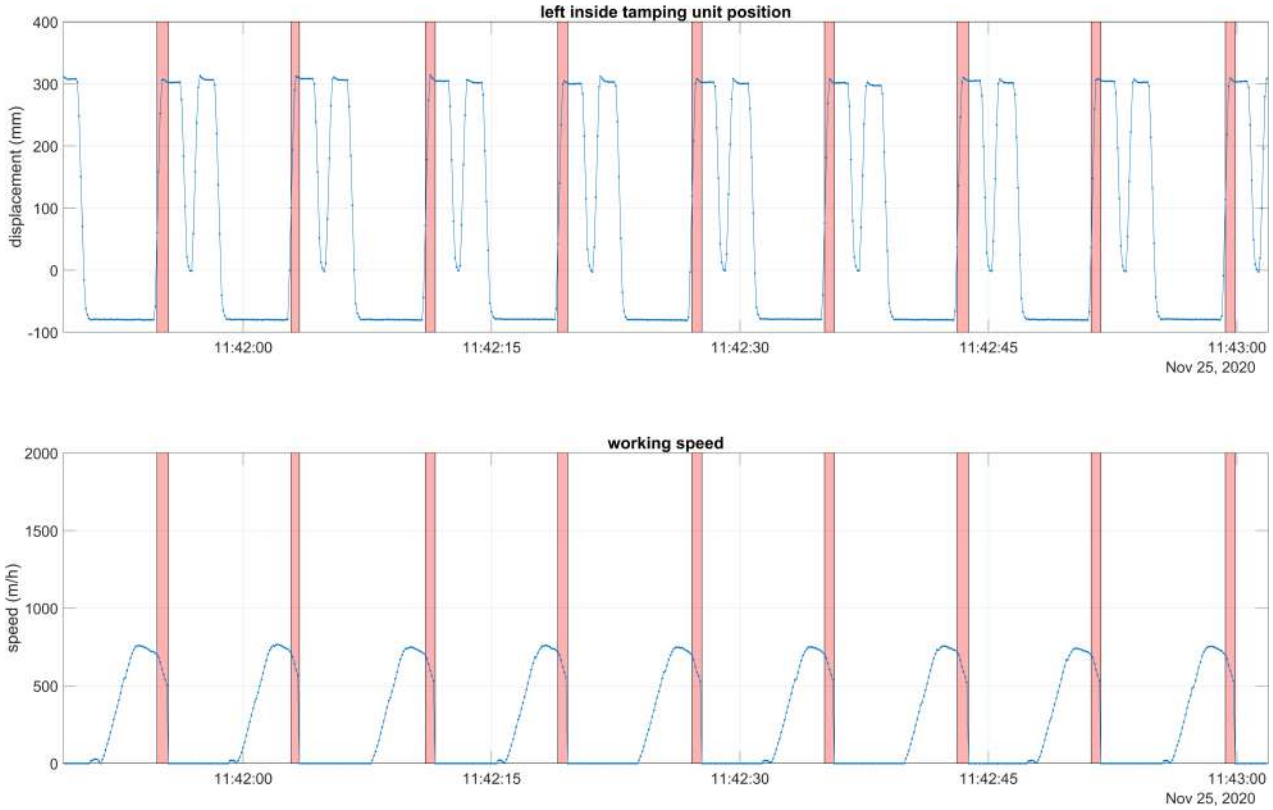


Figure 11. Identified NOK tamping cycle: The tamping unit is already in the ballast even though the machine is still moving

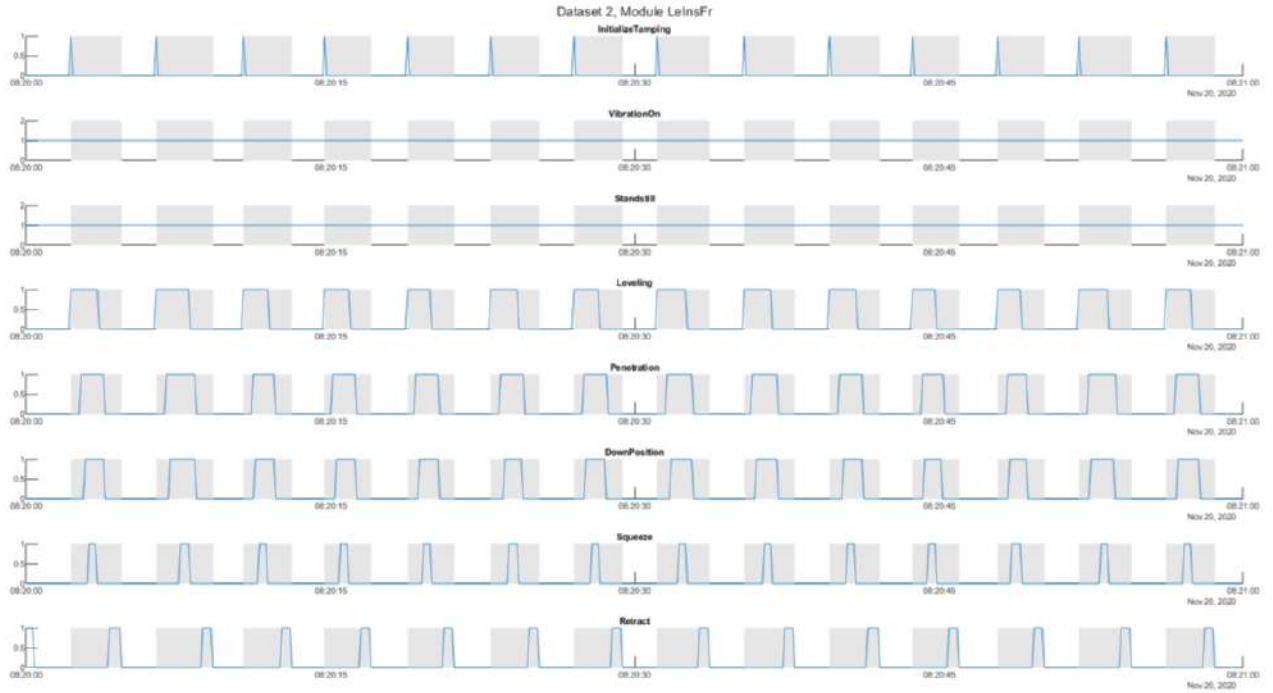


Figure 13. LSTM tamping cycle detection

Threshold Selection for Classification Models in Prognostics

Rohit Deo, Swarali Desai, Subhalakshmi Behera, Chetan Pulate, Aman Yadav and Nilesh Powar

Cummins Technologies India Pvt. Ltd., Pune, MH, 411045, India

Rohit.deo@cummins.com
Swarali.desai@cummins.com
Subhalakshmi.behera@cummins.com
Chetan.pulate@cummins.com
Aman.yadav@cummins.com
Nilesh.powar@cummins.com

ABSTRACT

In this study, we evaluate the performance of a prognostic classification model for NOX sensors in diesel engines over one month by comparing its predictions against actual outcomes. We then construct a validation dataset to assess the model's performance. By analyzing instances where the model's predictions were incorrect, we determine new threshold values that could potentially reduce errors for each false positive (FP) and false negative (FN). Subsequently, we create a dataset where the threshold varies for each observation and train a regression model with the modified threshold as the target variable. Our findings indicate that incorporating this approach, where the model's performance is iteratively refined using the validation dataset, leads to a reduction in both false positives and false negatives.

Keywords – True Negative (TN), True Positive (TP), False Negative (FN), False Positive (FP), Receiver Operating Characteristic (ROC), Area Under ROC Curve - (AUC)

1. INTRODUCTION

Cummins Inc. is a global corporation that designs, manufactures, and distributes engines, filtration, and power generation products. Cummins Inc. is headquartered in Columbus, Indiana, and has a history dating back to 1919. The company serves customers in more than 190 countries and territories, with a focus on innovation and sustainability in its products and operations. Prognostics plays a crucial role for Cummins in the context of diesel engines by enabling predictive maintenance. By analyzing the condition and performance of diesel engines using data from sensors and other sources, prognostics can help Cummins predict when maintenance or repairs will be needed.

This predictive approach allows Cummins to schedule maintenance in advance, minimizing downtime and reducing the risk of unexpected failures. Overall, prognostics help Cummins optimize the performance, reliability, and longevity of their diesel engines.

Diesel engines, the preferred power source for commercial vehicles like trucks and buses, produce harmful NO and NO2 emissions due to high combustion temperatures. Mckinley, Somwanshi, Bhave, and Verma, (2020) in their study showed that, to meet stringent emission standards, after-treatment systems such as selective catalytic reduction (SCR) are used, which can reduce emissions by factors of 10 to 20. SCR involves injecting a Diesel Exhaust Fluid (DEF) into the exhaust to produce ammonia (NH3), which then reacts with NOX to form harmless nitrogen (N2). NOX sensors are crucial in this process, measuring conversion efficiency and guiding the injection rate of DEF. Errors in these sensors can lead to either excessive ammonia or NOX emissions, impacting air quality and health. Regulatory agencies require continuous monitoring of these sensors and their operation to ensure compliance.

Prognostics aims to suggest changing out a NOX sensor before it fails. Since the replacement NOX Sensor can be planned for a convenient time rather than dealing with the discomfort of an unexpected breakdown, the customer will ideally experience less downtime. To determine whether the NOX sensor will fail, the present prognostics methodology uses a classification model with a predetermined threshold which is set using AUC ROC curve analysis. Even with an appropriate threshold, there may be instances where the model's predictions are not entirely accurate, which could potentially lead to increased downtime and maintenance costs for the customer. Dynamic thresholding model is a field of research that focuses on developing efficient methods for altering decision thresholds in predictive models over time and over different units.

The goal is to reduce the cost of false positives and negatives (after observing the performance of the prognostic classification model) by accounting for changes in the underlying data distributions that can arise because of changes in the environment, warranty status, or other external factors. In this study, we investigated a dynamic thresholding strategy and measured its performance by evaluating the incremental financial impact on customers and businesses. We also suggested a method for constructing a target label based on prognostics likelihood and validation data by computing the optimum thresholds. Our results demonstrate the importance of dynamic thresholding in maintaining the accuracy and robustness of predictive models and highlight the potential for further improvements through continued research in this area.

In summary, the paper explains how to dynamically change the threshold to improve the performance of a classification model after observing its performance for some time.

The rest of the paper is organized as follows. Section 2 gives a literature survey. Section 3 elaborates proposed Dynamic Thresholding Model (DTM) followed by results and discussions in section 4. Section 5 gives conclusions and future scope.

2. LITERATURE SURVEY

In this literature review, we will explore some of the key research papers that use ROC and AUC, dynamic thresholding, and cost analysis to determine thresholds.

Threshold selection is a crucial step in binary classification models as it determines the balance between the trade-off of precision and recall. Receiver operating characteristic (ROC) curves and area under the curve (AUC) are commonly used metrics to evaluate the performance of binary classification models and determine the optimal threshold value. The concept of the ROC (Receiver Operating Characteristic) curve, from which AUC ROC is derived, originated in electrical engineering and signal detection theory. It was initially used to analyze the performance of radar systems during World War II. The ROC curve was later adopted in medicine to evaluate diagnostic tests' performance. In machine learning, the ROC curve is used to assess the performance of binary classification models. The AUC ROC is a numerical measure derived from the ROC curve and provides a single value to quantify the overall performance of a classifier.

The ROC curve and AUC ROC can help in deciding the threshold for a binary classification model by providing insights into the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different threshold values. In their classic paper, Bradley (1997) emphasizes that AUC ROC provides a comprehensive measure of a classifier's performance across all possible thresholds, making it particularly useful for assessing the overall discriminatory ability of a model. The paper highlights that AUC ROC can be instrumental in threshold selection by illustrating the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at different threshold values. This insight enables practitioners to choose an optimal threshold based on the specific needs of the classification problem, balancing the costs associated with false positives and false negatives. In addition, Bradley discusses how AUC ROC can help in selecting a threshold that best suits the application's requirements. By analysing the ROC curve, which plots the true positive rate against the false positive rate at various thresholds, practitioners can visualize the classifier's performance and make informed decisions about threshold selection. This capability is particularly valuable in scenarios where the cost of false positives and false negatives differs, as it allows for the customization of the classifier's behaviour to meet specific needs.

Alotaibi and Flach (2021) introduce a novel approach to extend the traditional AUC metric to incorporate misclassification costs, addressing limitations in existing settings. By treating costs as sampled data, the proposed method employs the Weighted AUC (WAUC) metric and a novel estimator to approximate it, enabling a more accurate representation of model performance in complex cost-sensitive scenarios. The approach establishes a correspondence between WAUC and the cost function using threshold weighting and presents a bilevel optimization formulation to couple them. This formulation ensures that WAUC can be optimized at the optimal threshold value based on the real-world cost distribution. A stochastic algorithm is proposed for optimizing this formulation, demonstrating convergence rates comparable to standard SGD. Experimental results validate the effectiveness of the method in extending AUC to cost-sensitive scenarios, highlighting its significant performance improvements.

Yang, Yu, Wang, Quddus and Xue (2018) introduced the thresholding methods: fixed, rate-driven, optimal, RCut, MCut, and two novel ones: score-driven and global optimal are introduced. Score-driven thresholds can be adjusted globally or per label, offering flexibility. It investigates selecting a single global threshold or multiple thresholds. Using real-world datasets, the study conducts an empirical review, finding that the global and label-wise score-driven methods excel. Tuning a global threshold with respect to per-label cost is not significantly worse than using a separate threshold per label. Some traditional approaches, like the label-wise rate-driven method, may not suit highly imbalanced multi-label data. The study recommends using score-driven thresholds, globally or per label, for superior performance. It calls for further research on misclassification costs, loss, and threshold choice in multi-label classification, particularly when costs vary across labels.

In their study, Johnson and Khoshgoftaar (2019) showed that class imbalance is a common issue in machine learning, addressed through algorithm-level, data-level, and hybrid methods. While extensively studied in traditional algorithms, its application to deep neural networks (DNNs) is limited. This paper fills this gap by studying thresholding in DNNs using a Big Data Medicare fraud dataset. Employing random oversampling (ROS), random under-sampling (RUS), and a hybrid ROS-RUS, 15 training distributions with varying imbalance levels are created. Optimal classification thresholds are identified for each distribution on random validation sets, outperforming default thresholds. They further showed that, statistical analysis reveals a strong linear relationship between minority class size and optimal threshold, highlighting the importance of thresholding in DNNs for imbalanced data.

The properties of the F1 performance metric in multilabel classification, particularly regarding optimal decision-making thresholds. In this study, Lipton, Elkan, and Narayanswamy (2014) discuss how the best achievable F1 score is linked to the optimal threshold and highlights the impact of classifier behaviour in uninformative scenarios. For instance, in such scenarios, predicting all instances as positive maximizes the expectation of F1, which is beneficial for some metrics but problematic for others, like macro F1 in the presence of rare labels. The study also reveals that micro F1, on the other hand, maximizes the expected score by predicting all examples as negative in similar scenarios. This insight is especially valuable in settings with numerous labels. Additionally, the study suggests that micro F1 may wash out performance on rare labels. The findings underscore the importance of carefully selecting and understanding performance metrics, especially when choosing a single metric to optimize in scenarios involving competing systems, as this choice can significantly impact optimal thresholding behavior.

For a different application, Hancock, Johnson and Khoshgoftaar (2022) investigate the impact of the TPR $\geq$ TNR constraint on threshold values in classification tasks. The constraint favors lower thresholds, closer to the prior probability of the positive class, leading to reasonable trade-offs in classification rates. The default decision threshold of 0.5 is found unsuitable for the imbalanced Kaggle Credit Card Fraud Detection Dataset, yielding low TPR and FNR scores. It is noted that this default threshold is much larger than the prior probability of the positive class in imbalanced data. No single metric provides a comprehensive view of classifier performance, with thresholds closer to the positive class prior probability generally yielding better performance across multiple metrics. Each threshold selection technique offers trade-offs between positive and negative class performance. Starting with the positive class prior probability as a benchmark, thresholds can be adjusted to balance TPR and TNR scores. For specific performance goals, the optimal threshold can be estimated using the training dataset, considering user-defined performance metrics and constraints. The choice of performance metrics and constraints for threshold optimization significantly impacts test performance, highlighting the importance of careful selection based on the classification task's requirements and goals.

Going back in time, Chen, Tsai, Moon, Ahn, Young, and Chen (2006) explore the impact of decision thresholds on sensitivity, specificity, and concordance in four classification methods: logistic regression, classification tree, Fisher's linear discriminant analysis, and weighted k-nearest neighbour. While standard classification algorithms aim to maximize correct predictions (concordance), this may not be suitable for all applications. Some applications prioritize high sensitivity (e.g., clinical diagnostics), while others prioritize high specificity (e.g., epidemiology screening studies). The study examines the use of decision threshold adjustment to enhance sensitivity or specificity under specific conditions. Through Monte Carlo simulations, the study shows that increasing the decision threshold leads to decreased sensitivity and increased specificity, with concordance values remaining stable within an interval around the maximum concordance. Optimal decision thresholds can be identified within this interval to meet specified sensitivity and specificity requirements. The study analyzes three example datasets to illustrate these findings.

Two variants are introduced: a novel neural network-based thresholding method called ThresNets for improving multi-label predictions from class scores obtained from external scorers as shown by Shao & Huiyang et al. (2024) ThresNets are designed to scale linearly with the number of labels and can be trained offline after the scorer training is completed. One variant incorporates classic CS/CSS thresholds into the neural model, serving as a form of transfer learning between heterogeneous models. Our method is particularly suitable for medium-sized multi-label classification (MLC) tasks where informative label score dependencies can be found, and the ground truth of label assignments is reliable. Experimental results on artificially created scores demonstrate the effectiveness of ThresNets, especially when the scoring phase allows for improvements. ThresNets outperformed popular nearest neighbor-based classifiers in recovering from scoring errors. Empirical evaluation on real datasets shows that ThresNets perform better than classical methods according to various metrics, especially when used in a hybrid approach. Despite its advantages, ThresNets face challenges such as the risk of overfitting and the difficulty of training with long-tail labels. Future work will focus on leveraging external knowledge of class structure to improve ThresNets' performance further.

In their classic Real-time crash prediction, Draszawka, Karol and Szymanski (2023) show how important the Threshold selection is, by determining the cut-off point for

the posterior probability used to separate potential crash warnings. Current research lacks methods for effectively determining an optimal threshold, often resorting to subjective approaches. This study proposes a theoretical method using the mixed logit model to develop crash risk evaluation models. The minimum cross-entropy method outperforms other threshold selection methods, providing a reliable and automatic approach for identifying optimal thresholds in crash prediction.

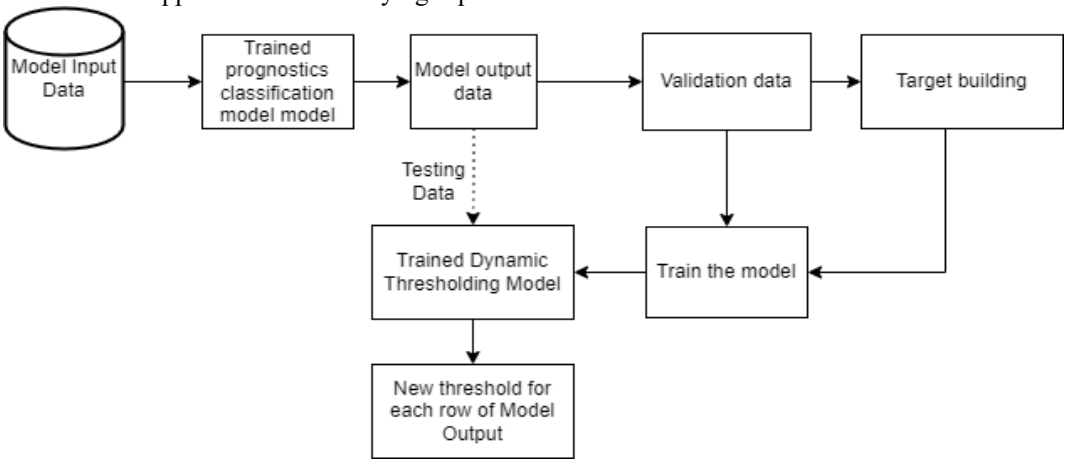


Figure 1. Block diagram of Dynamic Thresholding Model

In conclusion, this literature review has explored various aspects of threshold selection in machine learning, particularly in the context of class imbalance and multi-label classification. The reviewed papers have highlighted the importance of selecting an appropriate threshold for optimizing model performance and addressing specific challenges in different applications.

Several studies have proposed novel thresholding methods, such as ThresNets for multi-label classification and the use of the Area Under the ROC Curve (AUC ROC) metric for threshold selection. These methods have shown promising results in improving classification performance, especially in scenarios with imbalanced data and complex cost-sensitive considerations.

However, challenges such as overfitting and training with long-tail labels remain, suggesting the need for further research. Future studies could focus on leveraging external knowledge, such as class structure information, to enhance thresholding methods. Additionally, exploring the application of thresholding in emerging areas like real-time crash prediction and credit fraud detection could lead to valuable insights and advancements in the field.

3. PROPOSED METHODOLOGY

The prognostics model may make a future failure prediction for each component. We can run the prognostics model, determine the component's remaining useful life (RUL), and take preventive action. The model predictions for this classification model are based on the threshold and the likelihood for each data row. The threshold is often set using AUC-ROC curves, and that threshold is chosen for which the area under the curve for the TPR vs. FPR curve is highest. Our current prognostics algorithm alerts the client 90 days before the component's likely failure.

The dataset has two variations, pre-verified engine validation data, and post-verified engine validation data. Both have all engine status parameters, but they differ on when the actual status of the engine was captured. In the pre-verified engine validation data we predict engine health using a static threshold to classify an engine as healthy or faulty after we have the data about the actual status of the engine. In post-verified engine validation data, we predict before we have any ground truth for engine health using a static threshold to classify an engine as healthy or faulty, the ground truth is received later. The pre-verified engine validation data has been discarded due to data quality issues. The post-verified engine validation data has features such as the likelihood of an engine failure, the RUL of the engine, the classification threshold that was applied, the prediction by the predictive maintenance model, the actual status of the engine, and all other parameters of engine health.

Based on validation data in the suggested approach, we can choose the threshold dynamically. The validation data, with which, once the prognostics model has been run, we validate its performance during the following 90 days. Following the 90 days, we learn whether the failed event occurred. The confusion matrix, which we use to interpret the TPs, FPs, TNs, FNs, etc., is provided by the validation data. Our suggested methods can further reduce the FPs and FNs.

The order of events is depicted in the block diagram. The dynamic thresholding model is constructed utilizing features like odometer running, engine run duration, the likelihood of each row, the component's warranty status, and failure type (first or repeated failure), as opposed to setting a

threshold for the entire dataset. A regression model is the dynamic thresholding model. Concerning the validation data, the following algorithm is used to construct the regression model's target.

1. *For a TP and TN, there is no change in the threshold, fixed by the AUC-ROC curve.*
2. *E.g. - likelihood = 0.76, the fixed threshold is 0.8, which comes out to FN in the validation data, then, the new threshold is calculated as –*

$$Threshold_{New} = Threshold_{old} + |Threshold_{old} - Likelihood| - correction\_factor^{\#}$$

3. *Similarly, if likelihood = 0.86, the fixed threshold is 0.8, which comes out to FP in the validation data, then, the new threshold is calculated as –*

$$Threshold_{New} = Threshold_{old} + |Threshold_{old} - Likelihood| + correction\_factor^{\#}$$

4. *Repeat the procedure for each TP, FP, TN, and FN.*
5. # - The value can range between 0 to 1. This can be finalized after repeated training and testing of the model for different correction factors.

The target label, a continuous variable with a distinct threshold for each row of the output, is trained with the regressor model after we create the dataset containing the features as described before. After the model has been trained on the validation data, the DTM model that predicts dynamic threshold is given model out data with the same attributes as the validation data.

Extreme Gradient Boosting (XGBoost), a powerful machine learning method that we have used for this regression problem, is capable of handling complex non-linear interactions between features and targets as well as handling missing values and outliers.

The performance of the Xgb regressor used in DTM, the effect of changing the threshold, and the financial impact of this approach for each row are discussed in detail in the results and discussions.

4. Results and Discussions

The results (in Table 1) after applying Dynamic Threshold Modeling (DTM) show a notable improvement in various metrics. We have compared our results with traditional AUC ROC curves, as shown in column 2 of Table 1. We obtained the optimised threshold to be zero when we experimented with weighted AUC (WAUC) curves. This means that the model was recommending to replace every NOX sensor with even the smallest likelihood for failure. This is not a pragmatic solution. Also, The score-driven global thresholds proved to give no different results than the traditional AUC. Because of data availability constraints per the rest of the methodologies, we couldn't compare our methodology with them.

As compared to AUC, the True Positive Rate (TPR) has increased from 0.069 to 0.156, indicating that the model is better at correctly identifying positive instances. The Precision has also improved significantly, rising from 0.041 to 0.093, indicating a reduction in false positives. This improvement is further reflected in the F1 Score, which has increased from 0.074 to 0.14. Despite these improvements, the model still exhibits a relatively high False Positive Rate (FPR), albeit reduced from 0.397 to 0.373.

Table 1 shows the results for the EONOX sensor of a popular engine series. We observe that a reduction in FPs saves unnecessary repair of the engines and a reduction in FNs saves downtime cost of the engines which saves $0.3M for our customers and $13k for our company. In total, we save $313k. The methodology is highly scalable.

Overall, applying DTM has greatly enhanced the model's performance, particularly in correctly identifying positive instances and reducing false positives.

Table 1. Comparison of TPs, TNs, FPs and FNs before and after the use of DTM

	Using (AUC-ROC)	After DTM
TP	167	377
FP	3905	3670
FN	2246	2036
TN	5923	6158
TPR	0.069	0.156
FPR(or Recall)	0.397	0.373
Precision	0.041	0.093
F1 Score	0.074	0.14

5. Conclusions and Future Scope

The application of Dynamic Threshold Modeling (DTM) in our classification model has resulted in significant enhancements across key performance metrics, including accuracy, precision, and recall. By increasing the number of True Positives (TPs) and True Negatives (TNs) while decreasing False Positives (FPs) and False Negatives (FNs), the DTM has improved the model's overall effectiveness. We have implemented the DTM on one sensor from a specific family of engines used in a particular application. However, there is potential to expand this approach to multiple sensors across various engine families and to integrate it with other classification models. This scalability

could save substantial costs by minimizing unnecessary repairs and downtime. We estimate that such an expansion could result in savings amounting to millions of dollars.

REFERENCES

Mckinley, T., Somwanshi, M., Bhave, D. and Verma, S. 2020. *Identifying NOx Sensor Failure for Predictive Maintenance of Diesel Engines using Explainable AI*. PHM Society European Conference. 5, 1 (Jul. 2020), 11.

Andrew P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition, Volume 30, Issue 7, 1997, Pages 1145-1159, ISSN 0031-3203.

Reem Alotaibi, Peter Flach, *Multi-label thresholding for cost-sensitive classification*, Neurocomputing, Volume 436, 2021, Pages 232-247, ISSN 0925-2312.

Kui Yang, Rongjie Yu, Xuesong Wang, Mohammed Quddus, Lifang Xue, *How to determine an optimal threshold to classify real-time crash-prone traffic conditions?, Accident Analysis & Prevention*, Volume 117, 2018, Pages 250-261, ISSN 0001-4575.

J. M. Johnson and T. M. Khoshgoftaar, "Deep Learning and Thresholding with Class-Imbalanced Big Data," 2019, *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA, 2019, pp. 755-762.

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding classifiers to maximize F1 score*. arXiv preprint arXiv:1402.1892.

J. Hancock, J. M. Johnson and T. M. Khoshgoftaar, "A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data," 2022, *IEEE 8th International Conference on Collaboration and Internet Computing (CIC), Atlanta*, GA, USA, 2022, pp. 135-142.

Chen, J. J., Tsai, C. A., Moon, H., Ahn, H., Young, J. J., & Chen, C. H. (2006). *Decision threshold adjustment in class prediction*. SAR and QSAR in Environmental Research, 17(3), 337–352.

Shao, Huiyang, et al. "Weighted roc curve in cost space: Extending auc to cost-sensitive learning." *Advances in Neural Information Processing Systems 36* (2024).

Draszawka, Karol, and Julian Szymański. "From Scores to Predictions in Multi-Label Classification: Neural Thresholding Strategies." *Applied Sciences 13.13 (2023): 7591.*

BIOGRAPHIES

Rohit Deo received his Bachelor's degree in Electronics and Telecommunication from WIT, Solapur in 2011 and his Master's degree in Signal Processing from the University of Pune. He has eight years of professional experience. Rohit began his career as an Assistant Professor at Modern College of Engineering, Pune from July 2014 to May 2018. He then transitioned to industry, working as a Data Scientist at Montran Corporation, Mumbai from July 2018 to December 2021. Since December 2021, he has been a Senior Data Scientist at Cummins Technologies, Pune. His current and previous research interests include predictive analysis, pattern classification, machine learning, and deep learning for complex business problems. Rohit is highly skilled in signal processing and applied mathematics. He is a member of various professional societies.

Swarali Desai is a Data Science and Digital Engineering professional at Cummins. She holds a bachelor's degree in Electronics and Telecommunications Engineering from the University of Mumbai, India, and a Master of Science in Data Science from the University of Washington. With extensive experience in data analysis, machine learning, and business intelligence, Swarali has led several impactful projects in predictive maintenance, engine failure prediction, and supply chain optimization. Her recent work involves applying advanced data analytics techniques to enhance the performance and reliability of Cummins' engine systems, thereby contributing significantly to operational efficiency and customer satisfaction.

Subhalakshmi Behra holds a B.Tech in Mechanical Engineering from NIT Rourkela (2008) and an MBA in Business Management from XLRI Jamshedpur (2010). With over 14 years of industry experience, she currently serves as the Analytics Manager for the Analytics and Artificial Intelligence team at Cummins Inc. Her expertise spans advanced analytics, machine learning, digital product development, supply chain management, and project management. Subha is a Certified Six Sigma Green Belt and has numerous awards and honors for her contributions to the field.

Chetan Pulate, holds a Bachelor's degree in Electronics and Telecommunication from Smt. Kashibai Navale College of Engineering. With over 6 years of experience at Cummins Inc. as a Data Engineer, Chetan specializes in developing distributed, scalable, and reliable data pipelines. His expertise includes Apache Spark, Hive, Sqoop, Oozie, MapReduce, and various Hadoop stacks, along with working in Microsoft Azure. Chetan is skilled in feature engineering and deploying data science models at scale. He has also worked at Cognizant and Capgemini, enhancing his skills in big data and data engineering.

Aman Yadav holds a Bachelor of Technology degree in Computer Science from the University Institute of

Engineering and Technology, Kanpur (2021), and a Master of Technology degree in Artificial Intelligence from the Defense Institute of Advanced Technology (DIAT), Pune (2023). He is a Data Scientist at Cummins India, where he has been working since August 2023, after completing an internship as an M.Tech AI intern at the same company. Aman has experience in developing AI models, including an ASR model for Indian accents, gained during his internship at ProxMaq. His skills include machine learning, deep learning, NLP, and working with technologies like Apache Spark, Scikit-Learn, and Microsoft Azure. Aman is passionate about advancing AI and robotics, and holds several certifications in graph data science from Neo4j.

Nilesh Powar is the Advanced Analytics Director at Cummins. He holds bachelor in electronics engineering from University of Bombay, India, M.S in Computer Engineering from Wright State University and a doctoral degree in Electrical and Computer Engineering from University of Dayton, Ohio. He has over 20+ years' experience in field of image processing, machine learning, statistical pattern recognition and system integration. He had worked in the US as Distinguished Research Scientist for University of Dayton Research Institute, Dayton, OH, USA. Recent efforts involve data analytics for die casting, predictive analysis for supply chain management and video summarization using deep learning.

Design Of Digital Twins for In-Service Support and Maintenance

Atuahene Kwasi Barimah

Glasgow Caledonian University, Glasgow, G4 0BA, UK
abarim300@gcu.ac.uk

ABSTRACT

This research aims to examine the challenges in developing Prognostics and Health Management (PHM) analytics for Digital Twin (DT) use cases in industrial applications, with a particular focus on Multi-Component Degradation (MCD) scenarios. A hybrid methodology, integrating physics-informed and data-driven models, is employed, using limited asset degradation data for model development. Preliminary work includes an analysis of the impact of data quality on Fault Detection and Isolation (FDI) algorithm performance, as well as the proposal of a weighted ensemble hybrid approach for assets experiencing MCD scenarios Preliminary results indicate enhanced diagnostics in asset health management through the use of Physics-Informed models for FDI in MCD scenarios with limited prior degradation data. Expected contributions for this research are the development of physics-informed PHM analytics for DT applications in MCD scenarios, adaptive PHM analytics for evolving asset lifecycles in DT applications, and interpretable DT model analytics for PHM in systems facing Multi-Component Degradation.

1. BACKGROUND AND PROBLEM STATEMENT

Many high-value complex systems rely on advanced technologies, particularly the Industrial Internet of Things (IIoT), to monitor assets and carry out maintenance activities. Many stakeholders are increasingly turning to data-driven methods to monitor the condition of their assets, with Original Equipment Manufacturers (OEMs) offering various service packages in this regard (Barimah, Niculita, McGlinchey & Babakalli, 2021). These services leverage digital technology, particularly the concept of the digital twin (DT), to enable Prognostics and Health Management (PHM) applications. Digital twins serve as virtual replicas of physical assets (Grieves & Vickers, 2017), enabling operators to monitor, analyse, and predict asset states effectively. According to the Digital Twin consortium, the key capabilities required for digital twin use cases are data services, integration, user experience, intelligence, management and trustworthiness. This provides a framework for tailoring the capabilities of a digital twin for a particular industrial asset. The intelligence capability of a digital twin provides the requirements for enabling prognostic and health management applications. The analytics that drive intelligence in digital twins are constituted by either data-driven or knowledge-based models that provide insights for detection, diagnostics and prognostics for enhanced system reliability and support (Mihai, Yaqoob, Hung, Davis, Towakel, Raza, Karamanoglu, Barn, Shetve, Prasad, & Venkataraman, 2022).

However, developing the analytics that enable intelligence in digital twins (DT) for the full suite of PHM applications ─ detection, diagnostics and prognostics ─ is dogged with a lot of challenges which Compare, Baraldi and Zio (2019) describe in their work. One key challenge is the development of robust analytics for PHM applications, particularly with limited training data and in scenarios where assets are undergoing multi-component degradation (MCD) as part of a larger system. Bayesian approaches have been explored by Lin, Zakwan, and Jennions (2017) where the probability of two components in a fluid system was determined using a Bayesian probabilistic approach. However, there are limitations of this approach especially when compared with data-driven approaches in the MCD scenarios.

Another challenge in developing the analytics for PHM applications for Digital Twins is the evolution of the virtual replica of an asset throughout the asset's lifetime after commissioning and subsequent maintenance actions (Pires, Cachada, Barbosa, Moreira & Leitão, 2019), as various operating factors change the operating state of the asset, whether in a healthy or faulty condition. The complexity that the evolution of an asset throughout its lifecycle introduces in the DT development process presents model performance challenges for the PHM analytics embedded in the DT. Investigating how different DT model frameworks optimize analytics for PHM applications for an asset undergoing MCD scenarios will aid in identifying optimal DT model frameworks in the context of an evolving asset.

Lastly, the adoption of DTs has been increasing steadily in recent times, with data streaming and enhanced visualisation being some of the key selling points. However, the inadequate explainability in the outputs of the analytics capabilities limits the widespread adoption of DT analytics for critical assets (Presciuttini, Cantini, Costa & Portioli-Staudacher, 2024). In most DT applications, data-driven models that support the performance of DTs often train and perform as black boxes relying on the development of model weights which often become less intelligible (Kobayashi & Alam, 2024). Addressing the explainability of DT actions will facilitate the adoption of DTs for PHM applications, particularly for safety-critical systems.

2. EXPECTED CONTRIBUTIONS

This research aims to design a scalable online Hybrid-Digital Twin model architecture for IIoT-enabled PHM strategies for in-service support and maintenance applications. The expected contributions for this PhD research are presented below:

1. Development of physics-informed PHM analytics using limited training data for instantiating DT applications in Multi-Component Degradation scenarios.

2. Development of adaptive PHM analytics for evolving asset lifecycles in Digital Twin (DT) applications.

3. Development of interpretable DT model analytics for PHM applications in systems undergoing Multi-Component Degradation scenarios.

3. PROPOSED RESEARCH PLAN

The research plan seeks to advance the field of Hybrid-Digital Twins (H-DT) by defining an optimal physics-enabled model architecture for detecting and isolating Multiple Component Degradation (MCD) phenomena in complex systems. The research is based on a well-established testbed (see Appendix) to analyse the dynamic behaviour of a fuel system undergoing MCD scenarios. This hydraulic system comprises critical components, including a main supply tank, and an external gear pump driven by an induction motor. The rotational speed of the motor is regulated by a Variable Speed Drive (VSD). The system also features a solenoid shut-off valve (SHV) and five direct proportional valves (DPV1 to DPV5) for fluid flow control and fault emulation, respectively. Data collection is facilitated by pressure transmitters (P1, P2, P3, P4, and P5), turbine flow meters (F1 and F2), and a laser sensor to measure the pump's speed.

System components are connected using PVC tubing, and a finger valve is used for tank isolation when needed. In the context of fault simulation, specific control valves were manipulated to emulate fault conditions. For instance, DPV1 represented a clogged suction filter, fully open at 0% fault severity, while DPV2 simulated pump discharge side leakage and was fully closed at 0% fault severity. The SHV solenoid valve remained open, and DPV3, emulating a blocked or degraded shut-off valve, was fully open at 0% fault severity. DPV4 represented a clogged fuel nozzle, also fully open at 0% fault severity, while DPV5, simulating downstream pipe leakage, was fully closed with 0% fault severity. The healthy condition operating state of the system's control valves and their associated fault codes, as well as the test degradation scenarios for FDI Model Testing, are shown in the appendix.

The research will address several key objectives and research questions across four main areas:

1. Definition of a Hybrid-Digital Twin (H-DT): The research seeks to define an optimal architecture for H-DT models that meet diverse industry requirements and address MCD phenomena. This involves examining current trends and insights from literature via a systemic literature review.

2. Hybrid-Digital Twin Model Development: This phase involves the development of an online H-DT model tailored for MCD detection in complex systems. Key focus areas include data quality assessment, AI-enabled Fault Detection & Isolation (FDI), selection of hybrid model frameworks, and practical implementation considerations within Industrial Internet of Things (IIoT) systems.

3. System Reliability & Maintenance: The research will investigate the impact of employing an H-DT model architecture on different maintenance approaches, considering various technical, operational, and platform-specific requirements. The economic benefits of using H-DT models for different maintenance strategies will be analyzed.

4. Business Development: Finally, the research will focus on developing API-enabled services using a DevOps approach for the end-to-end implementation of the H-DT model architecture on IIoT platforms. It will explore the technical and business services enabled by an H-DT model architecture for cross-platform applications.

4. PROPOSED METHODOLOGY

The project will use the agile framework to plan various aspects of the digital twin development. A DevOps methodology will be adopted to facilitate the seamless integration and continuous deployment of physics-informed Prognostics and Health Management (PHM) analytics for the hybrid Digital Twin, focusing on components within the designated testbed. The industrial Internet of Things (IoT) platform Thingworx™ will be utilized to craft a user-centric experience (UX) for the digital twin, integrating the physics-informed PHM analytics for each asset through an Application Programming Interface (API) hosted on a remote server. Different physics-informed PHM analytics approaches will then be benchmarked on their performance in predicting multi-component degradation (MCD) scenarios in the context of challenges presented in the background of this report. Bayesian approaches will then be used to develop a trustworthiness framework to address the physics-informed PHM model uncertainty in predicting MCD phenomena. The Hybrid DT model testing and validation will be done using real-time data from the existing testbed and another testbed (proposed) which contains the same assets but in a different configuration. This will help in determining the scalability of the intelligence that underpins the predictive capabilities of the hybrid digital twin.

5. Research Work Done And Preliminary Results

5.1. Data Quality and FDI Model Performance

The relationship between data and predictive analytics was investigated in a recently published paper by Barimah, Niculita, McGlinchey and Cowell (2023). The analysis in this paper used data generated on the testbed described above as well as synthetic data to demonstrate high repeatability by the measurement system of the testbed. In the development of analytics for PHM applications, a lot of emphasis has been placed on data transformation for optimal model development without enough consideration for the repeatability of the measurement systems producing the data. This paper explored the relationship between data quality, defined as the measurement system analysis (MSA) process, and the performance of fault detection and isolation (FDI) algorithms within smart infrastructure systems using components of the testbed described above. The methodology employed starts with an MSA process for data quality evaluation and leads to the development and evaluation of fault detection and isolation (FDI) algorithms.

During the MSA phase, the repeatability of a water distribution system's measurement system was examined to characterize variations within the system. A data-quality process was defined to gauge data quality from the measurement system of the water distribution system. Synthetic data with varying data levels of quality levels was also used to investigate their impact on FDI algorithm development. Key findings reveal the complex relationship between data quality and FDI algorithm performance. The work carried out showed that synthetic data, even with lower quality, can improve the performance of a statistical process control (SPC) model, whereas data-driven approaches benefit from high-quality datasets. The study underscored the importance of customizing FDI algorithms based on data quality and a framework for instantiating the MSA process for IIoT applications, was also proposed for edge analytics which would be considered as part of future work.

5.2. Physics-Informed PHM for MCD scenarios

Optimizing PHM analytics for a system undergoing MCD scenarios using limited data was also investigated in the paper by (Barimah, Niculita, McGlinchey, Cowell and Milligan) and submitted to the PHME2024 conference which is currently under review. This study addresses the challenge of limited degradation data in developing Fault Detection and Isolation (FDI) models for multi-component degradation (MCD) scenarios. Utilizing a small fraction (1%) of the water distribution testbed dataset analyzed in the previous publication, a weighted ensemble hybrid approach was proposed and evaluated against more established modelling approaches. The proposed approach combines heuristic approximation and Physics-Informed Neural Network (Cai, Mao, Wang, Yin & Karniadakis, 2021) methods with a neural network model to enhance diagnostic performance.

The hybrid model generally outperforms other algorithms when tested on an MCD dataset, demonstrating improved diagnostic accuracy in such scenarios. This study contributes to the application of physics-informed FDI models for PHM applications in MCD scenarios, ultimately advancing asset health management. The paper also presents an ensemble FDI approach with the capability of addressing the limitations of integrating both data-driven and physics-based FDI models in multi-component degradation scenarios. Additional research will focus on dynamically optimizing ensemble hybrid model weights, leveraging prediction and model uncertainty to further enhance model performance for PHM applications.

6. Conclusion

In conclusion, this research endeavours to push the boundaries of Hybrid-Digital Twin (H-DT) technology, specifically targeting the challenges posed by Multiple Component Degradation (MCD) phenomena within complex systems. By researching issues of data quality assessment, fault detection and isolation (FDI) algorithm development, and the optimization of Predictive Health Management (PHM) analytics, some strides have been made. By studying data-driven and physics-based models, this research aims to propose a hybrid approach that optimizes diagnostic accuracy in MCD scenarios for PHM applications. The physics-informed PHM analytics developed from this research will improve on the current status quo by developing DT analytics models for PHM applications based on limited degradation data, adaptable in evolving asset lifecycles and intelligible. This will provide a new approach for addressing MCD scenarios aside from the use of classic Bayesian approaches for MCD prediction in the context of limited degradation data. This project will be relevant to industry because it will reduce the requirement for acquiring a lot of degradation data to train their degradation models ultimately reducing the cost of the FDI model development process for complex cases such as MCD scenario.

References

Barimah, A., Niculita, I.-O., McGlinchey, D., & Cowell, A. (2023). Data-quality assessment for digital twins targeting multi-component degradation in industrial Internet of things (IIoT)-enabled smart infrastructure systems. Applied Science, 13(24).

Barimah, A., Niculita, O., McGlinchey, D., & Alkali., B. (2021). Optimal Service Points (OSP) for PHM-enabled condition-based maintenance for oil and gas applications. 6th European Conference of the Prognostics and Health Management Society.

Cai, S., Mao, Z., Wang, Z., Yin, M., & Karniadakis, G. (2021). Physics-informed neural networks (PINNs) for

fluid mechanics. A review. Acta Mechanica Sinica, 1727-1738.

Compare, M., Baraldi, P. and Zio, E., 2019. Challenges to IoT-enabled predictive maintenance for industry 4.0. IEEE Internet of things journal, 7(5), pp.4585-4597.

Pires, F., Cachada, A., Barbosa, J., Moreira, A, P. and Leitão, P., "Digital Twin in Industry 4.0: Technologies, Applications and Challenges," 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 2019, pp. 721-726, doi: 10.1109/INDIN41052.2019.8972134.

Grieves, M. and Vickers, J., 2017. Digital twin: Mitigating unpredictable, undesirable emergent behaviour in complex systems. Transdisciplinary perspectives on complex systems: New findings and approaches, pp.85-113

Kobayashi, K. and Alam, S.B., 2024. Explainable, interpretable, and trustworthy AI for an intelligent digital twin: A case study on remaining useful life. Engineering Applications of Artificial Intelligence, 129, p.107620.

Lin, Y., Zakwan, S. and Jennions, I., 2017. A Bayesian approach to fault identification in the presence of multi-component degradation. International Journal of Prognostics and Health Management, 8(1).

Mihai, S., Yaqoob, M., Hung, D.V., Davis, W., Towakel, P., Raza, M., Karamanoglu, M., Barn, B., Shetve, D., Prasad, R.V. and Venkataraman, H., 2022. Digital twins: A survey on enabling technologies, challenges, trends and future prospects. IEEE Communications Surveys & Tutorials, 24(4), pp.2255-2291.

Presciuttini, A., Cantini, A., Costa, F. and Portioli-Staudacher, A., 2024. Machine learning applications on IoT data in manufacturing operations and their interpretability implications: A systematic literature review. Journal of Manufacturing Systems, 74, pp.477-486.

Development of a Data-driven Condition-Based Maintenance Methodology Framework for an Advanced Jet Trainer

Leonardo Baldo[1,2]

[1] *Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Torino, 10129, Italy*
leonardo.baldo@polito.it
[2] *Customer Services - Digital and Transformation Projects - Leonardo S.p.A. Aircraft Division, Caselle Torinese, 10072, Italy*
leonardo.baldo.ext@leonardo.com

ABSTRACT

Since their introduction more than 20 years ago, PHM strategies for aerospace equipment have gone a long way, enabling operators and Original Equipment Manufacturers (OEM) to monitor their assets, track down abnormal behaviors and plan maintenance action in advance. On the other hand, the transition from PHM strategies using simulated data to solutions utilizing real-life operational data is consistently prone to significant challenges and demands. This doctoral thesis aims to develop a PHM/CBM framework applied to a Electro-Hydraulic Actuators (EHAs) leveraging real in-service fleet data. In this paper, the first steps of the research project are presented.

1. INTRODUCTION

In the end of the 90s, the Joint Strike Fighter (JSF) Autonomic Logistics (AL) system began to take shape in the minds of forward-looking analysts and engineers with one mission: conceiving a revolutionary way to assist assets along their life cycle, hence enabling enlightened operational processes, innovative maintenance strategies and progressive logistic solutions (Smith, Schroeder, Navarro, & Haldeman, 1997; Hess & Fila, 2002). The AL framework core is encapsulated within Prognostic and Health Management (PHM) solutions which, as a consequence, have been defined as key enabling technologies for the development of reliable Performance Based Logistics (PBL) frameworks.
The creation of more available, dependable and resilient assets is especially important in the military aircraft sector, where the availability and reliability of assets are crucial for defense administrations to foster trust and guarantee mission readiness. Since the introduction of PHM strategies in the industrial and aerospace sector, in fact, many systems have been the scope of research in order to develop tailored prognostic strategies. It may then seem trivial that, along with other pivotal subsystems, the Flight Control System (FCS) is being gradually more covered by these approaches. However, this is only true to some extents.
While the constantly growing interest buildup involving the More Electric Aircraft (MEA) concept has led many prognostic research activities related to Electro-Mechanical Actuators (EMAs), applications on the widespread hydraulic actuators have somewhat lagged behind in terms of PHM. The challenges linked to the lack of precise and extensive data as well as the major difficulties in understanding and modeling failure mechanisms add one more difficulty layer to an already demanding task, which however deserves attention and can prove to generate extensive savings (Rodrigues, Yoneyama, & Nascimento Jr, 2012).

2. NOVELTY AND SIGNIFICANCE

The sharp contrast between the popularity of EHAs in both commercial and military aircraft and the scarcity of PHM related published studies focused on these actuators highlight a significant research gap - a gap that deserves attention.
The development of PHM solutions and strategies for such pivotal widespread systems holds substantial operational and economic potential for every stakeholder in the MRO sector. With the military MRO sector valued at around 37 billion USD in 2024, the demand for digital transformation initiatives and advanced MRO services is expected to undergo a substantial growth in the coming years, motivated by the necessity to maintain aging fleets and incorporate technological advancements for legacy equipment.
One way to address these performance requirements is focusing on the operations. The adoption of condition-based maintenance (CBM) and predictive maintenance (PDM) strategies falls within this enlightened vision which, thanks to the benefits offered by PHM analyses, provides decision makers with extended situational awareness of fleet operations. Some of the main components of a FCS are the actuators, which control the aerodynamic surfaces. Primary flight controls actuators are extremely safety critical elements within aircraft FCS

and, at the current state, they exploit mostly EHAs or Electro-Hydro-Static Actuators (EHSAs). Sure enough, found in most commercial and military aircraft, EHAs represent to date the backbone of actuation mechanisms for flight controls.

3. State of the Art

A solid literature base exists for individual actuator components. These studies have developed a wide range of solutions related to component-level fault detection and isolation (FDI), degradation models, and comprehensive PHM routines for individual parts: (Mi & Huang, 2023; Byington, Watson, & Edwards, 2004; Zhong et al., 2023) for servo valves, (Shanbhag, Meyer, Caspers, & Schlanbusch, 2021) for cylinders, (Vianna & Malere, 2014; Bertolino, Gentile, Jacazio, Marino, & Sorli, 2018) for leakages, (Chao, Shao, Liu, & Yang, 2023) for piston pumps. Additionally, some works have concentrated on the construction of custom test benches (Chiavaroli, De Martin, Evangelista, Jacazio, & Sorli, 2018) and the development of models (Iyaghigba, Petrunin, & Avdelidis, 2024), aiming at the generation of custom datasets. Moreover, some PHM strategies at the FCS level have been envisioned (Kosova & Unver, 2023; Shen & Zhao, 2023). Finally, most of the EHA level approaches found in literature primarily focus on the sole diagnostics (Iyaghigba, Ali, & Jennions, 2023). Notably, these approaches leave a consistent gap for EHA level PHM, which, to the best of the author knowledge, is approached with a limited number of strategies. In (Liu, Zhang, & Lu, 2015), the author developed EHA performance degradation predictions leveraging Elman neural network observer, support vector regression (SVR) and Gaussian Mixture Model (GMM). The research carried out in (Soudbakhsh & Annaswamy, 2017) and (Lu, Yuan, & Ma, 2018) shows the development of both a fault detection technique and a health monitoring approach. (Guo & Sui, 2020) presented an application of the Minimum Hellinger Distance on top of a Particle Filtering (PF). This PF-based solution is adopted by another PHM framework which combines also high-fidelity models (Autin, De Martin, Jacazio, Socheleau, & Vachtsevanos, 2021; De Martin, Jacazio, & Sorli, 2022). A modular hybrid fault prognosis method is developed in (Kordestani, Samadi, & Saif, 2020), where the author leveraged distributed neural networks and recursive Bayesian algorithm. In (Cui, Jing, Jiao, Huang, & Wang, 2023) the author approached a hybrid method: the nonlinear Wiener process (NWP) algorithm is used for the physics based section while the data-driven echo-state-network (ESN) is employed for the data driven one. In summary, the exhaustive yet limited number of studies mentioned above lay its roots on detailed actuator level data obtained from test benches and laboratory tests. Although highly valuable, the results of such studies hardly transfer to actual in-service legacy systems as detailed monitoring of low-level subsystem data is often not carried out and the control signals remain inside the FCC control loop without being saved or logged. On the other hand, the approaches that leverage operational data collected from real-world operational scenario are scarce and the few published studies provide constrained findings (Schoenmakers, 2020; Kannemans & Jentink, 2002).

In conclusion, if creating these frameworks was not an already challenging task, designing them for legacy and already operational platforms, definitely does not make the process easier. In this scenario, PHM engineers face obstacles related to working with pre-existing systems that were not originally designed for PHM applications (e.g low and/or variable sampling rates, limited built-in sensing/testing capability, no subsystem level sensors, hand written records, siloed databases, etc) as well as a vertical functional organization in the industry (Vogl, Weiss, & Donmez, 2014; Esperon-Miguez, John, & Jennions, 2013).

4. Approach and Work in Progress

This paper presents the initial steps towards implementing a comprehensive CBM framework for a specific aircraft subsystem. Precisely, the horizontal tail (HT) flight control Primary Actuation System (PAS) of an Advanced Jet Trainer (AJT), a twin-engine lead-in fighter training platform equipped with fully digital flight controls and avionics, is considered as a proof of concept (Baldo, De Martin, Sorli, & Terner, 2023). Through an in-depth analysis of design documents and operational procedures, relevant data have been identified and categorized. The AJT HT flight control PAS can be categorized as an EHA controlled by a tandem configuration Direct-Drive-Valve (DDV). The HT assembly is configured as an all-moving tail, a very popular solution when a good trade-off between control effectiveness, aerodynamic efficiency and operational complexity is desired. This solution has been adopted in various high-performance platforms (e.g. F16 Fighting Falcon, F22 Raptor) providing excellent maneuvering and flying qualities. On the other hand, DDVs are established solutions for flight controls and the adopted crank-connecting rod mechanism is widely accepted among mechanical solutions for longitudinal control.

The workflow employed for this research is reported in Figure 1. The most time demanding step so far has been represented by the domain understanding phase where the platform and data knowledge acquisition has been carried out. During this phase, significant effort has been devoted to acquiring comprehensive knowledge about the platform and gathering data. Leveraging the research group experience and expertise, both from the OEM and the University, the author created a data organization overview with the requested data for the first steps along with importance and priority indications. In this way, the author managed to reconstruct the data lineage and the data flow from the operative base to the info logistic systems and to the project data repository (DR). This first phase has been pivotal to plan ahead and understand which possible

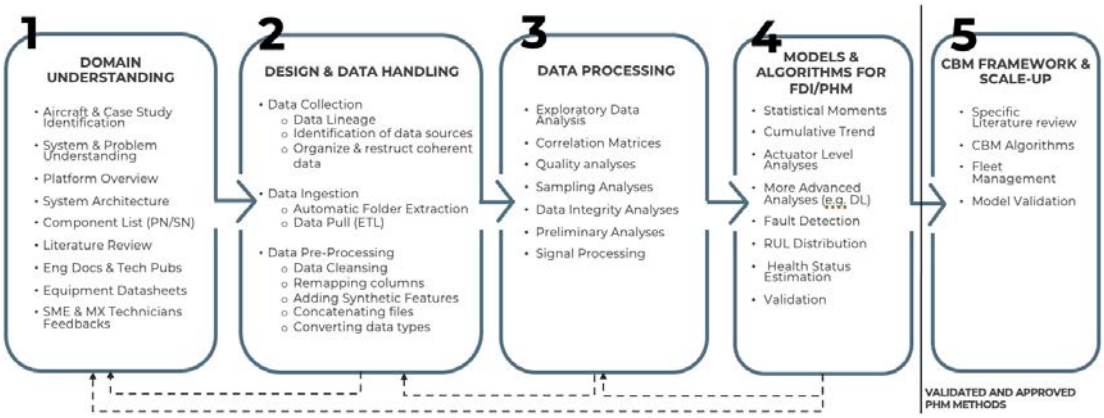


Figure 1. Research project workflow. Note the loopbacks to enhance the data processing with various iterations if necessary.

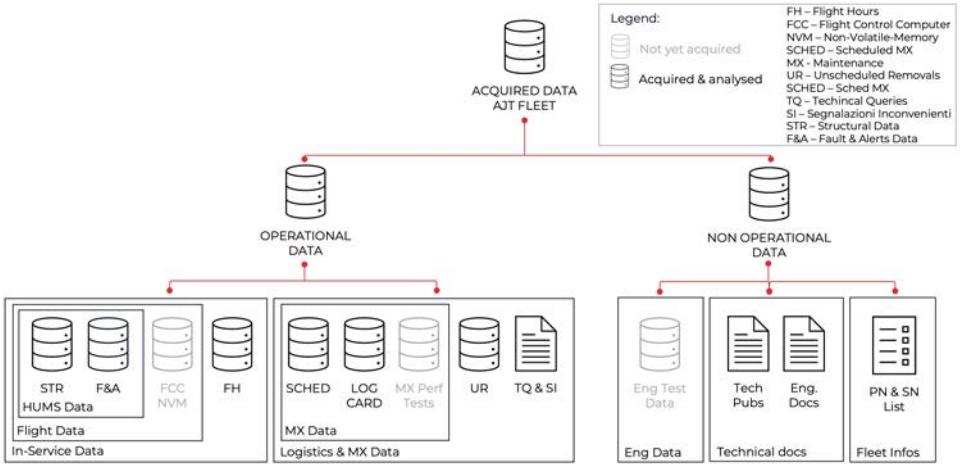


Figure 2. Data repository divided in categories.

strategic and methodological options could be approached, based on the available data. The current DR, encompassing more than 25000 Flight Hours (FH), is illustrated in Figure 2 where a clear distinction between operational (OD) and non-operational data (NOD) has been carried out to streamline the data classification process. LMX include Scheduled MX, unscheduled removals, log cards, technical queries, inconvenience reports and MX performance tests.

OD can be divided into In-Service Data (ISD) and Logistics & Maintenance Data (LMX). ISD includes all the data obtained from the aircraft itself after the sorties (FH register, Health Usage Monitoring System (HUMS) data and the Non-Volatile Memory of the FCC). In particular, HUMS data downloaded from the aircraft (S5000F, 2023) is divided in structural related data (STR) and Faults & Alerts (F&A).

Other potential operational data sources, which are often employed in the development of PHM strategies for legacy equipment, could include the Crash Survival Memory Unit (CSMU) or the Digital Video and Data Recorder. However, these latter sources were excluded from the study due to unreliable data download processes that occur only on an occurrence basis rather than consistently.

On the other hand, NOD encompass all technical information involving design, performance, process and configuration of aircraft components and subsystems (e.g. PAS PN and SN).

Following the domain understanding, the design and data handling and the data processing steps, the research is currently approaching the models and algorithm phase. This first steps focused on data derived from STR HUMS, Log Cards and UR. A total of 54 flight parameters (FP) has been selected through physical reasoning from the STR file. Given the lack of component-level signals that can accurately describe actuator health, relevant indicators were selected based on their potential to represent mechanical wear processes or possible flight anomalies (e.g. mechanical work). The selected FPs include:

- load components (forces and moments) acting on the HT and fuselage
- yaw, pitch and roll rates and accelerations
- body angles

- north, east, up speed components
- mobile surfaces deflections
- stick, pedals and throttle commands
- true air speed, Mach
- timestamps and complementary data
- 4 additional indirect signals (difference between two consecutive HT positions and the mechanical work carried out by the actuator obtained multiplying the position difference with the moment acting on the HT)

It is important to underscore once again that, by design, no actuator level data is recorded, including the actuator command produced by the FCC which would greatly benefit usage monitoring. FPs are saved in the form of time-series data with variable frequency. HUMS was not designed for PHM applications, thereby only a few irregular and sparse batches of high frequency data can be found in data records while most of the samplings are acquired at frequencies below 5 Hz. At the current state, this irregular low frequency sampling does not enable low level dynamic analyses of the actuator (whose dynamics is characterised by much higher frequencies) or the adoption of literature strategies based on high frequency actuator signals.

Following data quality and sampling analyses, the author thus decided to adopt a statistical approach based on cumulative features (CF). This approach has been chosen to determine if the data at hand demonstrates prognostic value in relation to the selected subsystem. CFs are currently being obtained from the merging of operational data sources and are the scope of current activities as reported in Figure 3. The four main statistical moments (SM) are calculated from the time series data of each flight for each FP. Then, CF are created by integrating these SMs in time (multiplying the FP SMs by the flight time) to replicate a time degradation tailored to the effective aircraft usage. Subsequently, the CF variations between two unscheduled removals are calculated and visualized using histograms representations. Histograms are then analysed and a signal-to-noise ratio ranking is performed to discern the most informative CF for further analyses and model development.

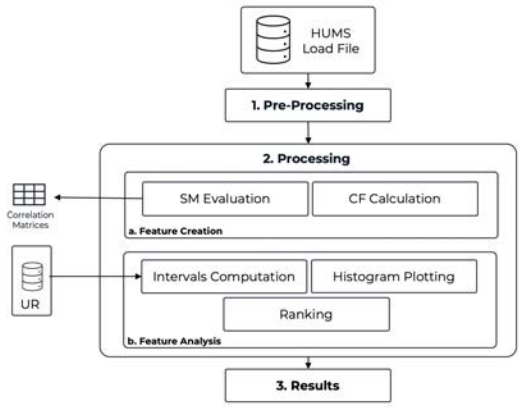


Figure 3. Statistical methodology overview.

The model and algorithm phase for diagnosis and PHM is currently being investigated and the calculations are currently being carried out. These results, if positive, would allow the author to statistically allocate a failure probability distribution in time leading to the next steps of the research project: conceiving a maintenance framework for fleet management leveraging a selected PHM strategy to support CBM. Otherwise, a custom actuator model will be needed to integrate in-service time series signals.

ACKNOWLEDGMENT

This publication is part of the project PNRR-NGEU which has received funding from the MUR – DM 352/2022. This research is cofunded by Leonardo SpA.

REFERENCES

Autin, S., De Martin, A., Jacazio, G., Socheleau, J., & Vachtsevanos, G. (2021). Results of a Feasibility Study of a Prognostic System for Electro-Hydraulic Flight Control Actuators. *International Journal of Prognostics and Health Management*, *12*. doi: 10.36001/ijphm.2021.v12i3.2935

Baldo, L., De Martin, A., Sorli, M., & Terner, M. (2023). Condition-based-maintenance for fleet management. In *Aerospace science and engineering - iii aerospace phd-days* (pp. 57–60). doi: https://doi.org/10.21741/9781644902677-9

Bertolino, A. C., Gentile, R., Jacazio, G., Marino, F., & Sorli, M. (2018). Ehsa primary flight controls seals wear degradation model. In *Asme international mechanical engineering congress and exposition* (Vol. 52002, p. V001T03A024). doi: 10.1115/IMECE2018-87080

Byington, C., Watson, M., & Edwards, D. (2004). Data-driven neural network methodology to remaining life predictions for aircraft actuator components. In *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No.04TH8720)* (Vol. 6, pp. 3581–3589). (ISSN: 1095-323X) doi: 10.1109/AERO.2004.1368175

Chao, Q., Shao, Y., Liu, C., & Yang, X. (2023). Health evaluation of axial piston pumps based on density weighted support vector data description. *Reliability Engineering & System Safety*, *237*, 109354. doi: 10.1016/j.ress.2023.109354

Chiavaroli, P., De Martin, A., Evangelista, G., Jacazio, G., & Sorli, M. (2018). Real time loading test rig for flight control actuators under phm experimentation. In *Asme international mechanical engineering congress and exposition* (Vol. 52002). doi: 10.1115/IMECE2018-86967

Cui, Z., Jing, B., Jiao, X., Huang, Y., & Wang, S. (2023). The Integrated-Servo-Actuator Degradation Prognosis Based on the Physical Model Combined With Data-Driven Approach. *IEEE Sensors Journal*, *23*(9), 9370–9381. doi: 10.1109/JSEN.2023.3248323

De Martin, A., Jacazio, G., & Sorli, M. (2022). Evaluation of Different PHM Strategies on the Performances of a Prognostic Framework for Electro-Hydraulic Actuators for Stability Control Augmentation Systems. In *Annual Conference of the PHM Society* (Vol. 14). (Number: 1) doi: https://doi.org/10.36001/phmconf.2022.v14i1.3289

Esperon-Miguez, M., John, P., & Jennions, I. K. (2013). A review of Integrated Vehicle Health Management tools for legacy platforms: Challenges and opportunities. *Progress in Aerospace Sciences*, *56*, 19–34. doi: https://doi.org/10.1016/j.paerosci.2012.04.003

Guo, R., & Sui, J. (2020). Remaining Useful Life Prognostics for the Electrohydraulic Servo Actuator Using Hellinger Distance-Based Particle Filter. *IEEE Transactions on Instrumentation and Measurement*, *69*(4), 1148–1158. doi: 10.1109/TIM.2019.2910919

Hess, A., & Fila, L. (2002). The Joint Strike Fighter (JSF) PHM concept: Potential impact on aging aircraft problems. In *Proceedings, IEEE Aerospace Conference* (Vol. 6, pp. 6–6).

Iyaghigba, S. D., Ali, F., & Jennions, I. K. (2023). A Review of Diagnostic Methods for Hydraulically Powered Flight Control Actuation Systems. *Machines*, *11*(2), 165. doi: 10.3390/machines11020165

Iyaghigba, S. D., Petrunin, I., & Avdelidis, N. P. (2024). Modeling a hydraulically powered flight control actuation system. *Applied Sciences*, *14*(3), 1206.

Kannemans, H., & Jentink, H. W. (2002). A Method to Derive the Usage of Hydraulic Actuators From Flight Data. In *Icas 2002 congress.*

Kordestani, M., Samadi, M. F., & Saif, M. (2020). A New Hybrid Fault Prognosis Method for MFS Systems Based on Distributed Neural Networks and Recursive Bayesian Algorithm. *IEEE Systems Journal*, *14*(4), 5407–5416. doi: 10.1109/JSYST.2020.2986162

Kosova, F., & Unver, H. O. (2023). A digital twin framework for aircraft hydraulic systems failure detection using machine learning techniques. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *237*(7), 1563–1580. (Publisher: IMECHE) doi: 10.1177/09544062221132697

Liu, H., Zhang, J., & Lu, C. (2015). Performance degradation prediction for a hydraulic servo system based on elman network observer and gmm–svr. *Applied Mathematical Modelling*, *39*(19), 5882–5895.

Lu, C., Yuan, H., & Ma, J. (2018). Fault detection, diagnosis, and performance assessment scheme for multiple redundancy aileron actuator. *Mechanical Systems and Signal Processing*, *113*, 199–221.

Mi, J., & Huang, G. (2023). Dynamic Prediction of Performance Degradation Characteristics of Direct-Drive Electro-Hydraulic Servo Valves. *Applied Sciences*, *13*(12), 7231. doi: 10.3390/app13127231

Rodrigues, L., Yoneyama, T., & Nascimento Jr, C. (2012). How aircraft operators can benefit from PHM techniques. In *Ieee aerospace conference proceedings.* doi: 10.1109/AERO.2012.6187376

S5000F. (2023). *S5000f - international specification for in-service data feedback* (Vol. Issue No. 3.1; Tech. Rep. No. S5000F-B6865-05000-00).

Schoenmakers, L. (2020). *Condition-based Maintenance for the RNLAF C-130H(-30) Hercules* (Unpublished doctoral dissertation). Eindhoven University of Technology.

Shanbhag, V. V., Meyer, T. J. J., Caspers, L. W., & Schlanbusch, R. (2021). Failure Monitoring and Predictive Maintenance of Hydraulic Cylinder—State-of-the-Art Review. *IEEE/ASME Transactions on Mechatronics*, *26*(6), 3087–3103. doi: 10.1109/TMECH.2021.3053173

Shen, K., & Zhao, D. (2023). A Fault Diagnosis Method under Data Imbalance Based on Generative Adversarial Network and Long Short-Term Memory Algorithms for Aircraft Hydraulic System. *Aerospace*, *10*(2), 164. doi: 10.3390/aerospace10020164

Smith, G., Schroeder, J., Navarro, S., & Haldeman, D. (1997). Development of a prognostics and health management capability for the Joint Strike Fighter. In *1997 IEEE Autotestcon Proceedings AUTOTESTCON '97. IEEE Systems Readiness Technology Conference* (pp. 676–682). doi: 10.1109/AUTEST.1997.643994

Soudbakhsh, D., & Annaswamy, A. M. (2017). Prognostics and health monitoring of electro-hydraulic systems. In *Dynamic systems and control conference* (Vol. 58288). doi: 10.1115/DSCC2017-5392

Vianna, W. O., & Malere, J. P. P. (2014). Aircraft hydraulic system leakage detection and servicing recommendations method. In *Annual conference of the phm society* (Vol. 6).

Vogl, G., Weiss, B., & Donmez, M. A. (2014). Standards for Prognostics and Health Management (PHM) Techniques within Manufacturing Operations. In *Annual conference of the phm society* (Vol. 6). doi: https://doi.org/10.36001/phmconf.2014.v6i1.2503

Zhong, Q., Xu, E., Shi, Y., Jia, T., Ren, Y., Yang, H., & Li, Y. (2023). Fault diagnosis of the hydraulic valve using a novel semi-supervised learning method based on multi-sensor information fusion. *Mechanical Systems and Signal Processing*, *189*, 110093. doi: 10.1016/j.ymssp.2022.110093

Digital Twin Development for Feed Drive Systems Condition Monitoring and Maintenance Planning

Himanshu Gupta[1], Pradeep Kundu[1]

[1]*KU Leuven, Brugge, 8200, Belgium*
himanshu.gupta@kuleuven.be
Pradeep.kundu@kuleuven.be

ABSTRACT

Current Prognosis and health management (PHM) technology suffers from challenges such as data availability, system interoperability, scalability, and transferability. In previous years, the PHM field has advanced a lot, but very few studies have been presented in which these challenges are addressed, and hence, PHM solutions are still confined to the lab environment. Digital Twin technology has the potential to address these challenges altogether and can add significant value to the PHM field. This thesis aims to develop an implementable Digital Twin framework for feed drive systems' condition monitoring and maintenance optimization, targeting these prevalent PHM challenges. The proposed framework will employ multiple physics-based models to generate synthetic data for different system states, configurations, and applications, and utilize this data with the help of machine learning to overcome the PHM challenges. The successful address of these challenges will pave the foundation in the direction of generalization of PHM solutions and also enhance the trustworthiness and reliability of PHM solutions.

Keywords-Digital Twin; Feed drive; Artificial intelligence

1. MOTIVATION

The feed drive systems primarily ball screw systems are used to convert rotary motion to linear motion and are employed in the field of manufacturing, machine tools, and robotics due to its high precision, and are used as electromechanical actuators for the aerospace and aviation components such as landing gear systems, flight control, engine actuation systems, etc.,(Qiao et al. 2018) where seamless and reliable operation is required. These systems have been employed due to high positioning accuracy and rigidity, which has been achieved by introducing preload between the screw and nut. Due to continuous operation, fatigue, and wear, these systems accumulate defects and lose preload over time, which leads to a loss in required precision and creates a backlash. Along with preload loss, the common fault modes for these systems are jam, spall, binding, and shaft bent (Yin et al. 2023). The mechanism failure of the feed drive is responsible for 18.72% of downtime in machine tools (Jia, Rong, and Huang 2019). As per the ASM handbook (Anon 1989) feed drive condition monitoring can decrease the production cost by up to 40% and increase the total productivity by 140% for machine tools.

For PHM of the feed drive, primarily physics-based or data-centric approaches are used (Butler et al. 2022). The physics-based approaches provide enhanced interpretability, but they are very sensitive to system parameters and fail to accommodate the uncertainties involved in the system. Data-centric approaches utilize the potential of machine learning. These approaches tend to be more accurate, but their accuracy depends on the amount of available historical data of the asset, which is not readily available in the field, making these approaches difficult to implement. Additionally, issues such as model interpretability may arise, hindering the understanding of how the model arrives at its conclusions.

Alternatively, digital twin (DT) technology offers a promising solution, which utilizes the concept of both physics-based and data-centric modeling strategies in synchronization and mitigates the shortcomings of both (K. Liu et al. 2022). Advancements in developing Digital Twins for feed drive systems have been relatively limited. W. Zhang et al. (2022) developed a DT framework for identifying rolling joints' dynamic parameters (stiffness and damping) for an FEA model. This involved conducting model tests on hardware and utilizing a DNN model in conjunction with the PSO algorithm to ascertain the parameter value. (D. Liu et al. 2022) developed a Digital Twin lumped mass dynamic model of a feed drive servo actuator system that maps the command and load information of the actuator to identify its vibration mode for the purpose of its health monitoring. (K. Liu et al. 2022) presented a multi-layer DT framework to predict and

compensate for time-varying positioning errors for a CNC machine. These existing DT frameworks focus on only one aspect of DT, i.e., developing virtual models, and fail to provide an implementable DT for a complete PHM solution with the capability of anomaly detection, diagnosis, and prognosis of the asset.

Additionally, existing PHM solutions face challenges like transferability, scalability, interoperability, and historical data availability. Transferability refers to the capability of employing a single PHM solution across various designs, materials, and configurations of the same asset. Scalability involves the adaptability of the PHM solution to the system used in diverse applications. Interoperability denotes the ability to apply the same PHM solution across different operating conditions of the asset. Lastly, historical data availability pertains to the accessibility and adequacy of data related to different fault severities and failures.

There have been progress in DT development for other applications which tries to tackle these challenges, such as (Feng et al. 2023) developed a DT framework for gear surface degradation monitoring. They utilized physics-based models to virtually represent the gear dynamics and employed optimization algorithms to fine-tune their dynamic parameters. These models were used to generate a data library for various degradation states, enabling transfer learning models to provide meaningful predictions by utilizing the vibration signal from the target asset. The framework is capable of adapting to uncertainties and demonstrates interoperability for different operating conditions. However, it focuses solely on monitoring specific defects, lacking a complete comprehensive solution. (Qi et al. 2024) Developed a DT-based monitoring system for the machining process of complex workpieces. The framework contains multiple layers and digital representation includes multiple models such as geometric, physics, behavior, and rule models. Real-time dynamic data is used for interaction mapping between the virtual models and the physical process. These tuned virtual models provide the state of the process. The framework applies to different operating conditions but fails to adapt to different applications and configurations.

This thesis aims to provide an approach for the development and implementation of a DT framework aimed at PHM of a feed drive system. The proposed framework aims to tackle the above defined PHM challenges.

2. PROPOSED METHODOLOGY

Figure 1. shows the proposed conceptual DT framework for condition monitoring and maintenance optimization, which has four distinct layers.

- Physical layer

The physical layer includes the monitored asset, with sensors for acquiring dynamic signals and necessary data acquisition devices. The layer provides the DT framework with the essential data for tuning the virtual model and for continuous monitoring of the asset.

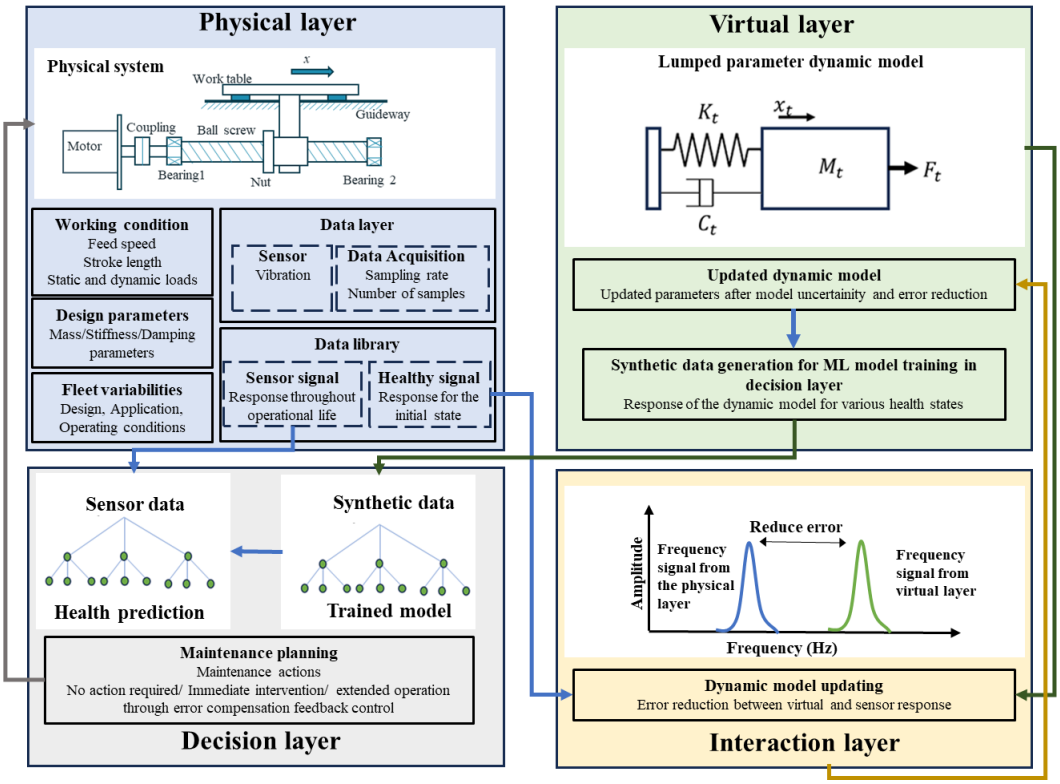


Figure 1. Proposed Digital Twin framework

- Virtual layer

The virtual layer mirrors the physical layer and simulates the dynamic behavior of the asset through either a lumped parameter dynamic model or an FEA model. The model employed is developed with adjustable parameters, which are adjusted as per the design configuration of the asset with the help of the interaction layer. This adaptability ensures that the virtual model correctly represents the physical asset.

- Interaction layer

This layer plays a crucial role in maintaining the virtual model's precise correspondence with the physical system. It accomplishes this by observing and modifying the virtual model's parameters to reduce any differences noted during comparisons of the model's behavior with the real-world responses of the physical system. The parameters will be tuned only once with the help of data for the healthy condition of the asset when the condition of the asset is known with relative certainty. This will help in overcoming the need for historical failure data.

- Decision layer

The tuned virtual model is utilized to create data repositories for various operational scenarios and fault severity stages of the physical asset. The decision layer utilizes this compiled data to train the machine learning (ML) models. After its training phase, the ML models employ the physical asset's responses as inputs to monitor the condition of the asset in real time. Further, the output from this health prediction model will be used for maintenance planning of the feed drive.

3. OBJECTIVES AND RESEARCH PLAN

The DT framework will be developed by focusing on each layer discussed in the previous section as a full objective. The proposed thesis has four objectives:

1. Configure a test rig with sensors and data acquisition hardware to gather data essential for the development and validation of the proposed framework.
2. Develop lumped parameter models for the test rig, enabling simulation of various faults, operating conditions, and applications of the feed drive. Additionally, create parametric and ML models to assess discrepancies between physical and virtual responses, facilitating adjustments to model parameters.
3. Develop ML models with the help of data from models created in objective 2 for health predictions.
4. Based on the output of the models from objective 3, develop maintenance planning strategies for the asset.

These objectives will be achieved by the following plan:

In objective 1, a test rig for the feed drive system will be configured. This rig would be used for multiple experiments related to different operating conditions and faults to closely emulate the environmental and operational variables encountered in an industrial setting. A triaxial accelerometer along with required data acquisition hardware will be used to gather vibration data from the rig. The vibration data will be further processed using advanced signal processing techniques such as wavelet transform and empirical mode decomposition to remove random noise components and extract the signal relevant to feed drive system dynamics only. The rig would be used to gather vibration data for three different nuts with different preloads to understand the effect of preload loss on the system's dynamics. The rig would later be used to collect data for the insinuated faults such as wear, nut and screw spalling, and backlash creating at least three datasets related to each fault. At last two run-to-failure experiments will be performed on the rig to understand the natural degradation of the feed drive. Based on the collected data health indicators will be selected for anomaly detection and diagnosis.

In objective 2, based on the configured test rig a lumped parameter model will be developed for the multiple rotational and translational degrees of freedom (DOF) of the feed drive. The model will incorporate the Hertzian contact theory to simulate the rolling elements and the Archard wear theory to simulate defects like wear. The model will be capable of simulating the experiment conditions planned with the rig. The models representing the interaction layer will be developed by using lumped models related to specific DOF and by using ML models that can provide a nonlinear mapping between the response of the physical asset and virtual model parameters such as stiffness and damping parameters. These parameters will be updated by comparing the difference in various features between the collected signal from the test rig and the synthetic signal generated through the virtual layer.

In objective 3, using the data gathered from objective 2 different explainable regression and classification ML models such as random forest, linear regression and decision tree models (Kundu, Darpe, and Kulkarni 2020) will be used for health assessment and damage quantification of the asset under the natural fault progression. These models will utilize the real-time vibration signal from the physical asset to predict its health state using features like natural frequency and ball pass frequency extracted from the signal. The predicted health state will further be utilized to estimate the stochastic positioning error. The initial concept of the proposed DT framework based on a single DOF model was developed and demonstrated in (Gupta and Kundu 2024).

In objective 4, the results obtained from objective 3 will serve as the basis for devising maintenance planning strategies for the asset based on the life cycle cost analysis. Further, the error compensation feedback control strategy will be formulated by utilizing an ML model to estimate required compensation based on location-specific positioning errors

estimated in objective 3. This is very important to utilize the maximum life of the feed drive.

Once all the objectives are fulfilled the DT framework will be ready to implement on any feed drive system. The framework designed will possess the flexibility to adjust parameters based on the specific asset and application, thus ensuring high transferability and scalability. Also, DT will be trained for different operating conditions and fault severities addressing interoperability and data availability issues. The current objectives of this thesis will focus on detection, diagnosis, and health management aspects.

Future work for this thesis could involve integrating the prognosis module and exploring the implementation of the framework on edge devices. Additionally, investigating order reduction techniques for both machine learning and physics-based models could help alleviate computing load and further improve the implementability of the framework.

4. CONCLUSION

This thesis aims to develop a digital twin framework for PHM applications that can provide better accuracy and versatility than physics-based and data-centric approaches. The proposed framework would be adaptable to different design configurations and also compensate for any system variations such as changes in operating conditions and applications. The research will contribute to tackling key PHM challenges such as transferability, scalability, interoperability, and historical data availability.

REFERENCES

Anon. 1989. *Machining*. ASM International.

Butler, Quade, Youssef Ziada, David Stephenson, and S. Andrew Gadsden. 2022. "Condition Monitoring of Machine Tool Feed Drives: A Review." *Journal of Manufacturing Science and Engineering* 144(10). doi: 10.1115/1.4054516.

Feng, Ke, J. C. Ji, Yongchao Zhang, Qing Ni, Zheng Liu, and Michael Beer. 2023. "Digital Twin-Driven Intelligent Assessment of Gear Surface Degradation." *Mechanical Systems and Signal Processing* 186:109896. doi: 10.1016/j.ymssp.2022.109896.

Gupta, Himanshu, Pradeep Kundu. 2024 "Digital Twin-driven Condition Monitoring: Development for Ball Screw Feed Drive Systems" Prognostics and System Health Management Conference, Stockholm, Sweden (May 2024), (Accepted)

Jia, Pingjia, Youmin Rong, and Yu Huang. 2019. "Condition Monitoring of the Feed Drive System of a Machine Tool Based on Long-Term Operational Modal Analysis." *International Journal of Machine Tools and Manufacture* 146:103454. doi: 10.1016/j.ijmachtools.2019.103454.

Liu, Dong, Shaoping Wang, Jian Shi, Shichang Qiao, and Guiling Liu. 2022. *Virtual Vibration Monitoring and Health Indicate for Ball Screw in Electromechanical Servo System: A Digital Twin Approach*.

Liu, Kuo, Lei Song, Wei Han, Yiming Cui, and Yongqing Wang. 2022. "Time-Varying Error Prediction and Compensation for Movement Axis of CNC Machine Tool Based on Digital Twin." *IEEE Transactions on Industrial Informatics* 18(1):109–18. doi: 10.1109/TII.2021.3073649.

Qiao, Guan, Geng Liu, Zhenghong Shi, Yawen Wang, Shangjun Ma, and Teik C. Lim. 2018. "A Review of Electromechanical Actuators for More/All Electric Aircraft Systems." *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 232(22):4128–51.

Yin, Zhengyang, Yi Yang, Guoji Shen, Yuehao Li, Liangyuan Huang, and Niaoqing Hu. 2023. "Dynamic Modeling, Analysis, and Experimental Study of Ball Screw Pairs with Nut Spalling Faults in Electromechanical Actuators." *Mechanical Systems and Signal Processing* 184. doi: 10.1016/j.ymssp.2022.109751.

Zhang, Wei, Di Zhu, Zhiwen Huang, Yidan Zhu, and Jianmin Zhu. 2022. "Dynamic Parameters Identification of Rolling Joints Based on the Digital Twin Dynamic Model of an Assembled Ball Screw Feed System." *Advances in Mechanical Engineering* 14(6). doi: 10.1177/16878132221108491.

Generating Realistic Failure Data for Predictive Maintenance: A Simulation and cGAN-based Methodology

Felix Waldhauser[1], Hamza Boukabache[2], Daniel Perrin[3], and Martin Dazer[4]

[1,2,3] *European Organization for Nuclear Research (CERN), Geneva, 1201, Switzerland*
felix.johannes.waldhauser@cern.ch
hamza.boukabache@cern.ch
daniel.perrin@cern.ch

[1,4] *Institute of Machine Components (IMA), University of Stuttgart, Stuttgart, 70569, Germany*
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

Absence of failure data is a common challenge for data-driven predictive maintenance, particularly in the context of new or highly reliable systems. This is especially problematic for system level failure prediction of analog electronics since failure characteristics depend on the actual system layout and thus might change with system upgrades. To address this challenge, this work pursues a novel simulation-assisted failure analysis methodology enabling automated and comprehensive evaluation of system level failure effects and failure detectability. While results obtained from simulations are suitable for comparative studies, they are confined to the simulation environment. To overcome this limitation, failure simulations are combined with generative models to generate realistic representations of missing failure data. Preliminary results demonstrate the capability of conditional generative adversarial networks (cGANs) to generate operational data of healthy systems, which accurately reflects correlations present in the source dataset. The proposed approach, using simulations as an additional source for generative models, not only targets the scarcity of failure data for highly reliable electronic systems but also ensures the adaptability of predictive maintenance algorithms to accommodate future system modifications and upgrades.

1. INTRODUCTION

Data-driven predictive maintenance of analog electronics requires algorithm-based detection of failure precursors in operational datasets containing voltage and/or current signals. However, obtaining sufficient historical failure data to train the algorithms, particularly for high-reliability or novel systems like safety instrumentation, is often a challenge. While manual studies of failure trajectories are feasible at the component level - such as examining discharge curves for capacitors - similar analyses at the system level involve studying numerous failure conditions, since failure characteristics not only depend on the components themselves, but also on their configuration in the system's layout. As a result, common failure trajectories may evolve with system upgrades or new generations, which would require validation or repetition of manual analyses. Hence, overcoming missing failure data requires a more automated approach, allowing exhaustive studies of failure characteristics while being adaptable to system upgrades.

The focus of this work is on developing a comprehensive simulation-assisted framework to establish a predictive maintenance algorithm for analog electronic systems in the absence of failure data. To illustrate this framework, a radiation monitoring electronics system, designed primarily for personnel safety, serves as demonstrator. It continuously monitors ambient dose rates and activates machine interlocks if defined radiation thresholds are exceeded. Given the system's critical role in ensuring safety, it is engineered to transition into a fail-safe mode upon detection of internal faults, initiating interlocks to mitigate risks. Thus, unforeseen failures triggering such interlocks can significantly impair the operational availability of downstream equipment. To address this conflict between safety and availability, the implementation of predictive maintenance based on data-driven failure prediction is proposed.

2. LITERATURE REVIEW AND RESEARCH CONTRIBUTIONS

Limited availability of failure data poses a challenge for failure prediction in industrial equipment, especially in systems

with high reliability and preventive maintenance (Rombach, Michau, & Fink, 2023), making failures observed during operation rare. Common approaches to overcome the scarcity of failure data include laboratory experiments to reveal failure modes (Janeliukstis, Ručevskis, & Kaewunruen, 2019), physics-based models incorporating relevant failure processes (Sun, Fan, Qian, & Zhang, 2016), or simulations to evaluate the system's response to failures (Mosleh, Montenegro, Alves Costa, & Calçada, 2021). In contrast, data-driven approaches, typically employing generative methods, eliminate the need for intricate system modeling. (Xiong, Fink, Zhou, & Ma, 2023) use generative adversarial networks (GANs) to extend already available failure data to new, unseen operating conditions based on a physics-informed loss function. While simulations of analog circuits are commonly employed for studies of failure effects (Zhang, Hong, Gao, & Yin, 2021), using simulations as a data source for generative models is currently limited to mechanical systems. (Gao, Liu, & Xiang, 2020) exemplify simulation-assisted data generation in the field of roller bearings, using FEM simulations to generate missing failure data via GANs.

This work explores the potential benefits of incorporating simulation-assisted failure analysis and data generation into predictive maintenance algorithms for analog electronics in the absence of failure data. The key contributions are:

- Utilization of synthetic datasets obtained from simulations to inform decision-making in the development of predictive maintenance algorithms at an early stage
- Generation of missing failure data using simulation-assisted generative methods bridging the gap between the healthy and the faulty domain
- Automated and resource-efficient framework for system level failure prediction in the absence of real failure data

3. Methodology and Preliminary Findings

In the frame of a feasibility study (Waldhauser, Boukabache, Perrin, & Dazer, 2022), unsupervised anomaly detection algorithms were applied to operational datasets of the radiation monitoring electronics system. The study demonstrated the capability of these algorithms to detect unusual data events, such as rare spikes of the dose rate measurement. Although the detected data events are technically anomalies, they are not necessarily related to hardware degradation or faulty behavior. Instead, they may represent atypical yet normal operational behavior. This results in the requirement of introducing knowledge on the system's failure behavior to establish the link between detected anomalies and the system's condition.

Since manual studies of the failure behavior are costly and not adaptable to design changes, alternative, more automated possibilities for acquiring comprehensive failure knowledge need to be explored. Here, simulations of the analog electronics using the SPICE simulation engine can be used to increase the understanding of the failure behavior. Specific failure scenarios are simulated by altering component characteristics, such as gradually reducing the capacitance of electrolytic capacitors. Hence, the impact of these failures on system-level outputs can be observed, facilitating the identification of failure patterns and assessing the detectability of component failures.

Additionally, simulation-derived datasets were used to identify the optimal source of failure knowledge for hybrid anomaly detection algorithms, which incorporate labeled failure data (Waldhauser, Boukabache, Dazer, Perrin, & Roesler, 2023). The results indicated that failure data derived from hardware tests, such as accelerated life tests, proved most beneficial in improving algorithm performance within this synthetic environment. Hence, the findings suggest prioritizing resources towards conducting hardware tests to gather failure data, as opposed to analysis of anomalous data events for failure identification by system experts.

4. Future Work and Research Strategy

The above mentioned studies emphasize the crucial need of understanding the system's failure behavior to refine failure prediction algorithms for identifying patterns indicative of hardware degradation. Failure simulations of analog electronics allowed detailed studies of failure detectability, common failure characteristics, and the generation of synthetic datasets. Although these simulations were suitable for comparative analyses, their utility is inherently limited to the simulation environment. Hence, future research endeavors will focus on bridging this gap between simulated and real datasets to ultimately compensate the missing failure data.

The subsequent phase of research therefore aims at manipulating operational datasets based on simulations to generate synthetic failure data. Here, one possible solution relies on generative artificial intelligence. Initially, a generative model is trained on data representing healthy system states with the objective of reproducing this baseline data. This methodology is then extended to address the generation of realistic failure data by fine-tuning the generative model with simulated failure scenarios without relying on real failure samples.

Preliminary studies with data of healthy states from radiation monitoring electronics containing measurements of internal voltages have demonstrated the capability of Wasserstein conditional generative adversarial networks (WCGANs) to generate synthetic data while preserving the correlations present in the training dataset. For example, WCGANs can accurately capture the relationship between temperature fluctuations and specific voltage signal characteristics. Figure 1 shows the comparison of real and WCGAN generated data for the kurtosis values of the 5 V signal.

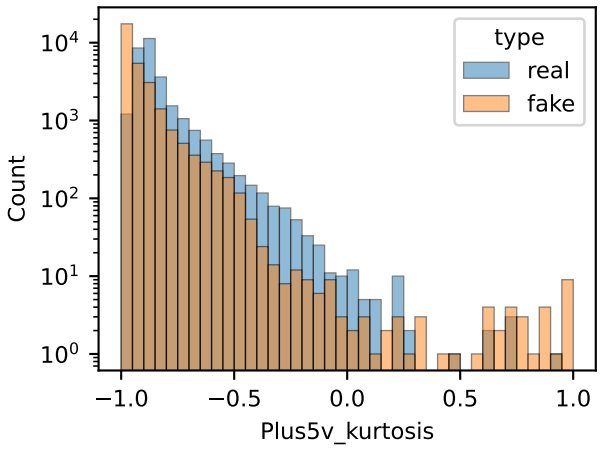


Figure 1. Histogram comparing the distributions of real and WCGAN generated (fake) data for the healthy system state on the basis of 30,000 samples per type. Plus5v_kurtosis is the kurtosis of the 5 V signal, with data normalized.

Ultimately, hardware tests of representative failure states will be required to validate the authenticity of the generated synthetic data. However, this necessitates accurate information regarding the health status of each component. Various options are being considered, including replacing components to replicate changes in their characteristics or inducing localized heat exposure to accelerate aging and confine failures to specific components.

5. Conclusion

The proposed methodology demonstrates a novel approach for developing predictive maintenance algorithms without relying on historical failure data. Simulations serve as additional knowledge source along the development process. This includes identifying detectable failures and generating synthetic failure data that is instrumental for training robust failure prediction algorithms. While hardware-based studies are still relevant, they are complemented by simulation results and limited to representative examples for validation purposes. Besides assisting the generation of missing failure data, comprehensive simulations of failure effects hold significant potential in automating analytical reliability assessments.

Acknowledgment

This work has been sponsored by the Wolfgang Gentner Programme of the German Federal Ministry of Education and Research (grant no. 13E18CHA).

References

Gao, Y., Liu, X., & Xiang, J. (2020, July). FEM Simulation-Based Generative Adversarial Networks to Detect Bearing Faults. *IEEE Transactions on Industrial Informatics*, *16*(7), 4961–4971.

Janeliukstis, R., Ručevskis, S., & Kaewunruen, S. (2019, November). Mode shape curvature squares method for crack detection in railway prestressed concrete sleepers. *Engineering Failure Analysis*, *105*, 386–401.

Mosleh, A., Montenegro, P., Alves Costa, P., & Calçada, R. (2021, December). An approach for wheel flat detection of railway train wheels using envelope spectrum analysis. *Structure and Infrastructure Engineering*, *17*(12), 1710–1729.

Rombach, K., Michau, G., & Fink, O. (2023, February). Controlled generation of unseen faults for Partial and Open-Partial domain adaptation. *Reliability Engineering & System Safety*, *230*, 108857.

Sun, B., Fan, X., Qian, C., & Zhang, G. (2016, November). PoF-Simulation-Assisted Reliability Prediction for Electrolytic Capacitor in LED Drivers. *IEEE Transactions on Industrial Electronics*, *63*(11), 6726–6735.

Waldhauser, F., Boukabache, H., Dazer, M., Perrin, D., & Roesler, S. (2023, October). Integrating System Knowledge in Unsupervised Anomaly Detection Algorithms for Simulation-Based Failure Prediction of Electronic Circuits. In *Proceedings of the 19th International Conference on Accelerator and Large Experimental Physics Control Systems* (pp. 249–256). JACoW Publishing.

Waldhauser, F., Boukabache, H., Perrin, D., & Dazer, M. (2022). Wavelet-based Noise Extraction for Anomaly Detection Applied to Safety-Critical Electronics at CERN. In *Proceedings of the 32nd European Safety and Reliability Conference* (pp. 1844–1851). Research Publishing, Singapore.

Xiong, J., Fink, O., Zhou, J., & Ma, Y. (2023, August). Controlled physics-informed data generation for deep learning-based remaining useful life prediction under unseen operation conditions. *Mechanical Systems and Signal Processing*, *197*, 110359.

Zhang, F., Hong, Z., Gao, T., & Yin, S. (2021, October). A Fault Detection Method for Analog Circuits Based on the Wavelet Features and One-class KNN. In *2021 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)* (pp. 1–6). Nanjing, China: IEEE.

Machinery Fault Detection using Advanced Machine Learning Techniques

Dhiraj Neupane, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal

School of Information Technology, Deakin University, Waurn Ponds, VIC 3216, Australia
{*d.neupane, reda.bouadjenek, richard.dazeley, sunil.aryal*} *@deakin.edu.au*

Abstract

Manufacturing industries are expanding rapidly, making it essential to detect early signs of machine faults for safety and productivity. With the extension of machines' runtime due to industrial automation, breakdown risks have increased, leading to economic and productivity consequences and sometimes even causalities. The surge in industrial big data from low-cost sensing technologies has enabled the development of intelligent data-driven Machinery Fault Detection (MFD) systems based on machine learning techniques in recent years. However, most existing methods are based on supervised pattern classification techniques to detect previously known fault types, which have limitations such as lack of generalization across different operational settings, focusing only on specific machinery and/or data types, and considering the identical and independent distribution of training and testing data. Therefore, my PhD research aims to develop a robust MFD framework for practical use by addressing these limitations.I will explore the potential of ensemble learning, unsupervised and semi-supervised anomaly detection, reinforcement learning, transfer learning, and cross-domain adaptation approaches in MFD. My PhD research will contribute to the field of data-driven MFD by proposing novel, effective solutions that can be applied across various manufacturing applications.

1. Background

Rotating machinery holds significant importance in modern industries. These machines often operate longer and under adverse conditions, making them prone to failure. Machine failures result in substantial maintenance costs, production inefficiencies, financial losses, and even risks to human life. Common electric motor failures involve bearings, stators, rotors, and gearboxes. The continuous operation of these machines can lead to wear, cracks, and other defects, emphasizing the need for accurate and timely fault detection and diagnosis to mitigate financial and safety risks (Neupane & Seok, 2020).

2. A Brief Discussion on the State-Of-The-Art

Recent developments (Neupane, Kim, & Seok, 2021; Zhong, Zhang, & Ban, 2023) in MFD have mainly focused on classifying the health states of machinery through extensive analysis of samples under normal and faulty conditions. While these studies have contributed to the creation of robust fault diagnosis systems, there is a limited exploration in examining semi-supervised learning (SSL) methods (see Figure 1). Moreover, the prior SSL applications primarily focus on fault classification (Zong et al., 2022; Zhang, Ye, Wang, & Habetler, 2020). Reinforcement Learning (RL) is increasingly being employed in various domains of MFD, such as transmission lines, hydraulic presses, and industrial process controls (Teimourzadeh, Moradzadeh, Shoaran, Mohammadi-Ivatloo, & Razzaghi, 2021; Junhuai, Yunwen, Huaijun, & Jiang, 2023). Although most RL applications treat fault diagnosis as a simple classification task, there are also some innovative approaches that extend its use to complex system management. For instance, (Vos, Peng, & Wang, 2023) employ an RL framework to optimize fleet management in the aviation sector, demonstrating how RL can effectively handle the dynamic decision-making required to maintain high fleet availability and minimize maintenance costs across aircraft with varying ages and degradation paths. Furthermore, data fusion methods play a critical role in enhancing the accuracy of fault detection systems. Techniques range from data-level fusion, such as weighted averaging and Kalman filters, to more complex feature and decision-level fusions that utilize statistical and machine learning methods, such as principal component analysis and Bayesian decision theory (Kibrete, Woldemichael, & Gebremedhen, 2024). Despite these advancements, the integration of multi-level fusion and cross-domain adaptation remains limited, highlighting a significant area for future research.

3. Motivations

First, the existing studies on MFD primarily employ supervised learning approaches (over 80%, see Figure 1), which can accurately identify known faults but struggle to detect

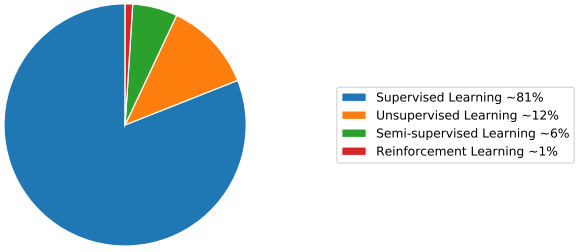


Figure 1. Machine learning techniques used for MFD

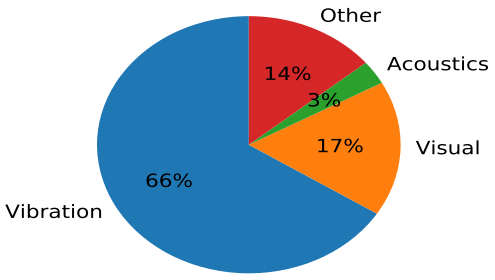


Figure 2. Data types used for MFD

novel or unseen ones (Das, Das, & Birant, 2023). This limitation is problematic in complex industrial settings where modern machines operate, as new fault types can emerge over time. Also, accurately identifying various fault types necessitates a substantial quantity of labelled data, which is a difficult challenge in real-world industrial settings because annotating takes time, expertise, and resources. Moreover, the labelled data may not cover all possible faults, restricting the diversity of the training dataset and hindering the model's ability to generalize to unseen faults.

In supervised learning, the effectiveness of the algorithms heavily relies on the accuracy of the data labels, which are typically derived from expert interpretations of sensor readings or from known operating conditions. This dependency on labeled data also extends to SSL, where limited labeled data can constrain the learning process to the inherent accuracy of these labels, a limitation referred to as the Bayes rate (Nian, Liu, & Huang, 2020). To address this challenge, RL can be used, which is a promising ML framework that learns from trial-and-error interactions using rewards rather than explicit instructions (Wang, Jiang, Li, & Liu, 2020). RL has demonstrated success in various fields, including manufacturing, but its application in MFD is limited. Existing fault detection systems have not thoroughly exploited the potential of RL in optimizing maintenance decisions and fault detection strategies. Current RL algorithms for MFD often treat fault diagnosis as a simple classification task, which may not fully utilize the capabilities of the RL framework. RL can learn the sequence of events leading up to a fault, which can be used to predict when a fault is likely to occur. Moreover, the use of RL algorithms is currently limited to a single machinery or environmental setting. This research aims to explore the application of RL algorithms in MFD and develop specialized RL algorithms that can effectively handle the dynamic nature of fault patterns and machine operational conditions.

Additionally, most existing (about two-thirds, see Figure 2) ML-based MFD techniques use vibration data to predict faults (Das et al., 2023). Vibration signals, however, can be problematic in harsh environments or areas with high background noise, which can decrease the accuracy of collected data. Moreover, traditional vibration sensors may not always be practical for installation in locations that are difficult to access or on specific types of equipment. For example, placing these sensors on ball bearings within centrifugal pumps or on equipment operating under extreme conditions such as low-temperature vacuum pumps can pose significant challenges (Hoang & Kang, 2019). Furthermore, an exclusive reliance on vibration data could potentially limit the performance of ML models. Thus, the incorporation of other data types could offer a richer understanding of the problem and yield improved results. Utilizing diverse data types, such as temperature, current, acoustic, and visual information, into fault diagnosis algorithms can offer a more comprehensive and accurate understanding of machinery health. Combining data from multiple sources not only improves the detection of subtle faults, reducing diagnostic errors but also compensates for potential sensor failures or data inaccuracies due to environmental interference.

Moreover, most existing work on MFD focused on specific machine types and operational environments. Usually, models are trained on data from one type of machine in a particular environment and are expected to perform effectively on similar machines or the same machine under different conditions. This expectation is based on the assumption that the source (training) and target (test) data are independent and identically distributed (*iid*) (Li et al., 2022). However, achieving this *iid* condition in industrial applications is challenging due to several factors: (a) machines can exhibit different behaviors and degradation patterns over time or when operated in varying conditions; (b) differences in machine manufacturing, wear-and-tear, and operational settings can introduce significant variability in the data. These factors contribute to the 'domain-shift' problem, where the training data no longer represents the new conditions under which the model is tested. This domain shift can significantly reduce the effectiveness of fault detection models, as they fail to generalize across different operational scenarios. Thus, addressing this issue is crucial for affecting machine health monitoring and fault diagnosis in diverse environments.

4. Research Aim and Objectives

This PhD project aims to develop a robust framework for MFD by addressing identified limitations and research gaps. Defined as 'robust', our framework ensures that various algorithms—including unsupervised, semi-supervised, and reinforcement learning techniques—-perform effectively in real-world industrial settings. These environments are often complex and noisy, with heterogeneous data. By utilizing diverse data types, our framework anticipates supporting effective applications across different domains. This aim will be achieved through the following objectives:

1. To investigate the potential of unsupervised and semi-supervised anomaly detection (AD) methods for identifying anomalous patterns in MFD. This approach eliminates the need for labeled data and addresses the challenges of class imbalance.
2. Additionally, there is a goal to fully utilize the capabilities of RL for MFD by creating specialized algorithms that can optimize maintenance and fault detection strategies. Apart from fault classification, RL has potential in AD (Arshad et al., 2022), optimizing maintenance strategies (Marugán, 2023) or prediction (Siraskar, Kumar, Patil, Bongale, & Kotecha, 2023). This will help overcome the limitations of treating fault diagnosis as a guessing game and improve performance in diverse operating conditions.
3. Another objective is to explore the potential of using diverse data types for developing MFD algorithms, which can enhance diagnostic efficiency and provide a comprehensive understanding of machinery health status. The study will also investigate the use of ensemble models to improve accuracy and efficiency.
4. Lastly, the aspiration is to bridge the gap between different data types and operational settings using domain adaptation and transfer learning techniques, which can enhance the model's ability to generalize across diverse settings.

5. Research Methodology and Timeline

To create an integrated framework for MFD that makes use of robust semi and unsupervised learning-based AD algorithms, our approach encompasses data preprocessing, algorithm selection, and model training, with the aim of generating anomaly scores for predicting faults. We will evaluate the performance of our models using metrics such as precision, recall, F1 score, etc., and compare them with supervised methods. Our work on this project is ongoing, and we submitted an article to "the 8th European Conference of the PHM Society (PHMe2024), presenting the results of our experiments with various AD algorithms on the Case Western Reserve University (CWRU) bearing dataset, Paderborn University (PU) bearing dataset, and Health and Usage Monitoring System (HUMS) datasets. The outcomes of our study so far have been encouraging, demonstrating the efficacy of AD methods.

To achieve our second objective in employing RL in MFD, the formulation of problems, the development of algorithms (including state representation, action space definition, reward function design, and RL algorithm selection), data collection and preprocessing, training and testing, and continuous refinement of the RL algorithm based on evaluation metrics such as performance against baseline models, rewards evaluation, fault detection accuracy, and training convergence progress will be done. Since real-time data collection is limited in our setup, we will focus on employing offline RL techniques (Deng, Sierla, Sun, & Vyatkin, 2023). Offline RL is ideal for situations where learning must be derived from pre-existing datasets rather than from interactions with the environment in real-time. For implementing these techniques, we can utilize well-established libraries, which provide the necessary tools to effectively apply offline RL algorithms to our data.

To accomplish our third objective, we aim to develop a comprehensive MFD algorithm by integrating various data types. Our aim is to improve adaptability, generalization, accuracy, and fault detection capabilities under different machinery conditions. We employ flexible models that can handle heterogeneous data, which are preprocessed for noise and normalization, and utilize ensemble techniques like data, feature, or decision fusion. Evaluation will be based on accuracy, precision, recall, and F1 score metrics. We have made progress by using the PU dataset to fuse vibration and current data, which will gradually advance to the integration of X and Y-axis vibration data, two phases current data, and torque data, ensuring comprehensive feature integration and decision-making.

To effectively enhance MFD in diverse operating conditions and overcome the challenges of limited data by employing domain adaptation and transfer learning techniques, we will apply domain adaptation methods like discrepancy, adversarial, or reconstruction-based approaches (Zhang et al., 2023). Moreover, multi-source domain adaptation is also being explored. The effectiveness of the approach is evaluated using classification and domain discrepancy metrics.

5.1. Expected Outcomes and Publications

This project aims to develop a novel, robust, and flexible MFD system for real-world applications. The framework will implement an unsupervised or semi-supervised learning-based fault detection framework, utilizing diverse data types and incorporating RL for fault prediction and cross-domain platforms as well. Apart from these, we have expected to publish a review paper, which is almost ready to submit. Also, a few conference articles and collaboratory publications are also expected.

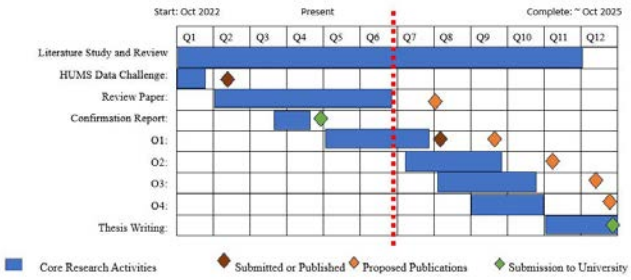


Figure 3. Proposed Timeline for this PhD

5.2. Works Done So Far and Timeline

During the course of this research project, significant progress has been made across various aspects of the study.

Literature Review and Draft Article: Our initial phase included a comprehensive literature review and a draft of a detailed review paper on significant MFD developments. The paper covers multiple topics and is almost ready for submission.

Participation in HUMS Data Challenge and Conference Presentation: We participated in the HUMS Data Challenge and presented our research at the HUMS conference, where our findings were published. Our work involved using signal processing and statistical-based approaches to detect cracks in the provided dataset and the auto-regressive integrated moving average method for predicting fault progression.

Manuscript Submission in PHMe2024: We also submitted an abstract and full manuscript for the upcoming conference PHMe2024, going to be held on July 3- 5, in Prague, CZ. The article is related to the use of semi-supervised-based techniques for machinery fault detection, which is objective 1 of this project.

Objective 1 continued: The work being carried out for objective 1 is being extended. We are actively engaged in the incorporation of a new dataset and in the application of semi-supervised techniques for anomaly detection. Furthermore, we are exploring various data transformation methods and combinations of features to enhance our results.

This PhD program commenced in October 2022 and is expected to be completed by 2025, within a three-year timeframe. In addition to core research and publication activities, administrative tasks must be carried out throughout the program, as per university and faculty regulations. To facilitate effective planning, the entire three-year period has been divided into twelve three-month periods and is shown in figure 3.

6. Conclusion

Implementing a robust machinery fault detection system in real-world settings presents several challenges, such as adaptability to a variety of machines, compatibility with existing infrastructure, and scalability across diverse industrial environments. To tackle these challenges, we aim to develop adaptive algorithms, enhancing system compatibility with current technologies, and ensuring scalability for broad industrial applications. Moving forward, our vision for MFD research encompasses the integration with predictive analytics, aiming to transform MFD systems into comprehensive diagnostic tools that not only detect but also predict faults, significantly reducing downtime and maintenance costs. This future-oriented approach aims to solidify the role of MFD in advancing predictive maintenance strategies and thereby contribute a sustained impact to the field.

References

Arshad, K., Ali, R. F., Muneer, A., Aziz, I. A., Naseer, S., Khan, N. S., & Taib, S. M. (2022). Deep reinforcement learning for anomaly detection: A systematic review. *IEEE Access*, *10*, 124017–124035.

Das, O., Das, D. B., & Birant, D. (2023). Machine learning for fault analysis in rotating machinery: A comprehensive review. *Heliyon*.

Deng, J., Sierla, S., Sun, J., & Vyatkin, V. (2023). Offline reinforcement learning for industrial process control: A case study from steel industry. *Information Sciences*, *632*, 221–231.

Hoang, D. T., & Kang, H. J. (2019). A motor current signal-based bearing fault diagnosis using deep learning and information fusion. *IEEE Transactions on Instrumentation and Measurement*, *69*(6), 3325–3333.

Junhuai, L., Yunwen, W., Huaijun, W., & Jiang, X. (2023). Fault detection method based on adversarial reinforcement learning. *Frontiers in Computer Science*, *4*, 1007665.

Kibrete, F., Woldemichael, D. E., & Gebremedhen, H. S. (2024). Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review. *Measurement*, 114658.

Li, W., Huang, R., Li, J., Liao, Y., Chen, Z., He, G., . . . Gryllias, K. (2022). A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges. *Mechanical Systems and Signal Processing*, *167*, 108487.

Marugán, A. P. (2023). Applications of reinforcement learning for maintenance of engineering systems: A review. *Advances in Engineering Software*, *183*, 103487.

Neupane, D., Kim, Y., & Seok, J. (2021). Bearing fault detection using scalogram and switchable normalization-based cnn (sn-cnn). *IEEE Access*, *9*, 88151-88166. doi: 10.1109/ACCESS.2021.3089698

Neupane, D., & Seok, J. (2020). Bearing fault detection and diagnosis using case western reserve

Natural Language Processing for Risk, Resilience, and Reliability

Jean Meunier-Pion[1]

[1]*CentraleSupélec Université Paris-Saclay, Gif-sur-Yvette, Essonne, 91190, France*
jean.meunier-pion@centralesupelec.fr

ABSTRACT

Natural Language Processing (NLP) has seen a surge in recent years, especially with the introduction of transformer architectures, relying on the now famous self-attention mechanism. Especially, with the rise of Large Language Models (LLM), propelled by the appearance of ChatGPT in 2022, a new hope of extracting relevant information from text has emerged. In the meantime, natural language data have not often been used in risk, resilience, and reliability tasks. However, text data containing reliability-related information, that can be used to monitor health information regarding complex systems, are available in several and diverse shapes. Indeed, text data can either contain theoretical expert knowledge (technical reports, documentation, Failure Modes and Effects Analysis (FMEA)), or in-practice expert knowledge (incident reports, maintenance work orders), or in-practice non-expert knowledge (customer feedback, news articles). Critical infrastructures, such as nuclear powerplants, railway networks, or electrical power grids, are complex systems for which any failure would induce severe consequences affecting many people. Such systems have the advantage of serving many users, thus having many possible text sources from which technical information and past incident data can be mined for anticipating future failures and generating responses to catastrophic scenarios. The goal of this work is to develop methods and apply state-of-the-art NLP techniques to text data relating to critical infrastructures and failures, to (1) mine information from unstructured language data, and (2) structure the extracted information. Preliminary experiments were conducted on customer review data and incident reports, and show promising performance for failure detection from text data with transformers, as well as incident-related information extraction using LLMs.

1. STATEMENT OF THE PROBLEM ADDRESSED

Risk, resilience, and reliability have seen some attempts to use Natural Language Processing (NLP) to make use of text data in systems health monitoring. NLP was applied to maintenance records data so as to filter maintenance records by types (Stenström et al., 2015). Sharp et al. (2017) also developed a framework on maintenance records, to classify such data based on expert tags and by supervised learning. Considering the specificity of technical terms used in maintenance work orders, Brundage et al. (2021) introduced the notion of Technical Language Processing (TLP) and discussed the need for models designed and trained specifically on technical language data. Other works (Li & Wu, 2018; Huang et al., 2021) have proposed a statistical approach to look at co-occurrences of terms and a graph visualization to quickly perceive how failures are characterized in diesel engines, based on Failure Modes and Analysis Effects (FMEA) data. Research in NLP for risk, resilience, and reliability covers multiple applications, with different datasets and tasks. However, it suffers from a lack of common shared open-source datasets and benchmarks, and with the rise of generative artificial intelligence, there is currently room for improving existing frameworks and developing new ones.

The research question addressed in this thesis is the following: how can one extract information from unstructured text data, and then structure the extracted information, to learn failure knowledge from text data?

The initial approach should involve using state-of-the-art NLP techniques, especially Large Language Models (LLMs) and transformers (Vaswani et al., 2017) in general, to extract information from text. The extracted information will then be organized in knowledge databases, according to ontologies, in order to structure the information relating to risk, resilience, and reliability. The goal is to use the large amount of available text data containing health information of complex systems so as to learn and structure knowledge on failures of critical infrastructures.

To that end, various forms of text can be used. Either documents containing theoretical expert knowledge, such as technical reports, technical documentation, or FMEA; or documents with in-practice expert knowledge, such as incident reports, or maintenance work orders; or documents with in-practice non-expert knowledge, such as customer feedback, or news articles. Such data can then be used in two complementary ways: either to directly mine information

from them, or to give context and knowledge while extracting information from other documents.

The expected benefits are the creation of tools in the form of specialized technical search engines and automated text assistants to support informed decision-making for the anticipation of incidents and the generation of response scenarios when encountering failures in critical infrastructures.

2. Expected Contributions to the Field

The main expected contributions to the field include (1) the development of open-source datasets to support NLP tasks applied to risk, resilience, and reliability, (2) the application of state-of-the-art NLP techniques, including LLMs to reliability data and the creation of associated benchmarks, (3) the design of an ontology for reliability engineering and the development of a method to automatically populate knowledge databases whose architecture would rely on this ontology.

3. Research Plan

The research plan currently includes the following parts: (1) detecting failures and assessing reliability from text data, (2) applying LLMs for information extraction, (3) focusing on failure mode extraction with the proposed framework for information extraction assisted by LLMs, (4) designing an ontology for reliability engineering, and (5) automatically populating knowledge databases for system reliability.

As a transversal task, the development of fine-tuned LLMs for risk, resilience, and reliability tasks, e.g., including code generation for reliability engineering, is a common thread.

3.1. Failure Detection and Reliability Assessment from Text Data

The simplest unit of information that can be extracted from text data regarding reliability is whether or not the document at hand states that a failure occurred.

Following previous research (Meunier-Pion et al., 2021), a set of customer review data for failure detection was developed for the task of detecting if customers report a failure in their review of a product. It is composed of 2,415 customer reviews labeled for binary classification. Additionally, labels include a level of granularity that enables the subtask of classifying failures severity as tolerable or intolerable.

Due to the ambiguity of customer reviews, several annotators were required to label the dataset and a human benchmark score was derived from the annotations to know what the best performance of a machine model could be. The human performance was estimated to 91.24% of balanced accuracy, while the best model involving a fine-tuned DeBERTa-v3 transformer (He et al., 2023) reached 88.50% balanced accuracy. This constitutes promising results for detecting failures in customer review data, and in natural language in general, in order to generate lifetime data from text corpora and assess reliability directly from natural language data.

The results from this research part suggest that reliability-related information can be extracted from text data. Building upon this preliminary work, the aim of this thesis is to gather more fine-grained information regarding systems health, such as failure causes, failure modes, degradation, maintenance actions, interdependencies between system components, and so on.

3.2. Application of LLMs to Information Extraction

With the rise of LLMs and their incredible capabilities for understanding natural language, it seems that NLP information extraction tasks can be addressed more effectively. However, one limitation of LLMs is that they are designed for generating text, in the form of long consecutive sentences, instead of returning only a specific word or set of words answering a short query.

In this research, LLMs were applied on nuclear powerplants incident reports data for extracting basic information such as the date of an incident and the place of an incident. Using a small LLM stored on less than 3 GB, an average accuracy of 94.5% could be reached for the extraction of date and place of incidents, over an initial dataset of 50 incident reports. Besides, one should note that if a LLM outputs “The date of the incident was 2023.”, then the output is considered invalid, as the expected queried information is only “2023”, making the task more challenging as conciseness matters.

The goal is to provide a framework for extracting information thanks to LLMs, that combines the ability of LLMs to understand text and generate high quality answers, with a methodology for extracting specific queried information. Here, in this part of the research, the goal is not necessarily to extract technical information, but rather to come up with an effective and performant framework for extracting pre-defined attributes when queried, as illustrated in Figure 1.

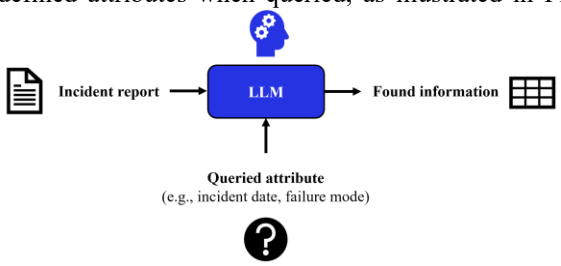


Figure 1. Information extraction using an LLM.

3.3. Failure Mode Extraction

By leveraging the framework developed for Section 3.2., one type of reliability-related information that can be queried

from text is the failure mode of a given system, i.e., how the system failed. As part of this research, methods are developed to extract failure mode information from text.

There are mainly two ways of extracting failure mode information from text: (1) classifying failure mode based on pre-defined failure mode labels, or (2) generating a failure mode label that fits the given description of an incident.

While the first way, involving multi-class classification may be convenient to assess performance and compare models, it requires the definition of class labels, which are not always available, especially when working with new unseen data. On the other hand, the second envisioned approach to extracting failure mode from text involves generating labels, which requires a more sophisticated evaluation framework. This approach is intended in this research, in order to leverage LLMs for information extraction. Additionally, one modern technique that will be investigated in this research is Retrieval-Augmented Generation (RAG), which consists of generating an answer to a query, with the addition of a context from a vector database and similar to the input query.

In the meantime, the first way of extracting failure mode information is currently under study and preliminary results on the National Highway Traffic Safety Administration (NHTSA) complaints dataset show that it is possible to reach 86% balanced accuracy on multi-class classification of failure modes on text data, using only standard NLP techniques, without even the use of transformers.

3.4. Definition of an Ontology for Reliability

The objective of this research being to learn failure knowledge from text data, one important part of this work is to define an ontology for reliability. Previous works in maintenance have already applied ontology frameworks to define ontologies like an Ontology model for Maintenance Strategy Selection and Assessment (OMSSA) (Montero Jiménez et al., 2023).

The purpose of defining an ontology for reliability is to organize concepts relating to failures in order to structure failure knowledge. This should enable and facilitate the automatic instantiation of knowledge databases containing failure information extracted from text data.

3.5. Automatic Population of Knowledge Databases

Ultimately, the purpose of this research is to enable the automatic population of knowledge databases containing failure-related information extracted from text data.

In that respect, a challenge that will be addressed in this research is grouping fields of the same data record. Indeed, multiple data records can have their information in the same document and an additional challenge thus is: how to distinguish between different data records? How can one group fields together to create the correct instance, and not mix fields from different records together?

More specifically, fields that can be extracted from text data include, for example, the date of an incident, the failure mode, and the root cause of the failure. The challenge is to correctly map the date of incident A with the failure mode and root cause of A, and not map it with the failure mode and root cause of B, whenever A and B co-occur in a document.

3.6. Fine-Tuning LLMs for Reliability Engineering

As part of this research, a transversal component will be the development of fine-tuned LLMs specialized on technical data in order to efficiently use technical engineering data and to address tasks relating to system health monitoring.

In that respect, a first attempt of benchmarking LLMs on the fields of risk, resilience, and reliability, is under study and involves the creation of a dataset for code generation containing more than 50 code generation questions. This dataset is inspired by the HumanEval dataset (Chen et al., 2021) and involves the usage of unit tests to guarantee the capability of the model to generate effective code. The goal is to evaluate current state-of-the-art LLMs, such as variations of Mistral or Llama models, on the vertical application of risk, resilience, and reliability, whereas traditional code generation benchmarks (Austin et al., 2021; Du et al., 2023) consist of general programming tasks.

Then, an LLM will be fine-tuned on specific data to compensate for the lack of expert knowledge from general LLMs, and enable the generation of more accurate technical scripts from an artificial intelligence code assistant. This approach will be generalized to fine-tune LLMs not only for code generation, but also for natural language in general, in order to acquire expert knowledge on complex systems.

4. CONCLUSION

The current research aims at developing open-source datasets and benchmarks for NLP for risk, resilience, and reliability, while leveraging state-of-the-art techniques like LLMs. The main focus here is the development of methods for information extraction and structuring knowledge.

Preliminary results show encouraging evidence that state-of-the-art NLP techniques are able to mine failure-related information from text data. Nonetheless, the methods developed in this thesis are intended to be applied to critical infrastructures, thus confidence indicators are necessary to measure the trustworthiness of the developed models.

As a common thread, an objective throughout this research is to create NLP-related materials, including datasets and code, that will be shared to encourage research in this field and ensure access to trustful and reproducible results.

REFERENCES

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). *Program Synthesis with Large Language Models.* Preprint. Google Research.

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, vol. 27, pp. 42-46. doi: 10.1016/j.mfglet.2020.11.001

Chen, M., et al. (2021). *Evaluating Large Language Models Trained on Code.* Preprint. OpenAI, San Francisco, California, USA.

Du, X., Liu, M., Wang, K., Wang, H., Liu, J., Chen, Y., Feng, J., Sha, C., Peng, X., & Lou Y. (2021). *ClassEval A Manually-Crafted Benchmark for Evaluating LLMs on Class-level Code Generation.* Preprint. Fudan University, Shanghai, China.

He, P., Gao, J., & Chen, W. (2023). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.* Preprint. Microsoft Azure AI.

Huang, Q., Wu, G., & Li, Z. T. (2021). Design for Reliability Through Text Mining and Optimal Product Verification and Validation Planning. *IEEE Transactions on Reliability*, vol. 70, pp. 231-247. doi: 10.1109/TR.2019.2938151

Li, Z., & Wu, J. (2018), A Text Mining based Reliability Analysis Method in Design Failure Mode and Effect Analysis. *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*. June. doi: 10.1109/ICPHM.2018.8448909

Meunier-Pion, J., Zeng, Z., & Liu, J. (2021). Big Data Analytics for Reputational Reliability Assessment Using Customer Review Data. *Proceedings of the 31st European Safety and Reliability Conference (ESREL 2021)*. September 19-23, Angers, France. pp.2336-2343

Montero Jiménez, J. J., Vingerhoeds, R., Grabot, B., & Schwartz, S. (2023). An ontology model for maintenance strategy selection and assessment. *Journal of Intelligent Manufacturing*, vol. 34, pp. 1369-1387. doi: 10.1007/s10845-021-01855-3

Sharp, M., Sexton, T., & Brundage, M. P. (2017). Towards Semi-autonomous Information: Extraction for Unstructured Maintenance Data in Root Cause Analysis. *IFIP International Conference on Advances in Production Management Systems*. March 9, London, England. pp.425-432

Stenström, C., Aljumaili, M., & Parida, A. (2015). Natural Language Processing of Maintenance Records Data. *International Journal of Condition Monitoring and Diagnostic Engineering Management*, vol. 18, pp. 33-37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017), Attention is All you Need. *Advances in Neural Information Processing Systems*.

Prognostics of Remaining Useful Life for Aviation Structures Considering Imperfect Repairs

Mariana Salinas-Camus[1], Nick Eleftheroglou[2], and Dimitrios Zarouchas[3]

[1,2] *Intelligent Sustainable Prognostics Group, Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands*
m.salinascamus@tudelft.nl
n.eleftheroglou@tudelft.nl

[3] *Center of Excellence in Artificial Intelligence for Structures, Aerospace Engineering Faculty, Delft University of Technology, Delft, The Netherlands*
d.zarouchas@tudelft.nl

ABSTRACT

Maintenance plays an important role in fulfilling the goals of the Prognostics and Health Management (PHM) field. As of now, no publication has addressed the impact of imperfect repair actions from the prognostics perspective. Imperfect repairs introduce complexities, altering system degradation processes and increasing prediction uncertainties, thereby impacting the accuracy of Remaining Useful Life (RUL) predictions. To fill this gap in the literature, the study proposes developing a robust prognostic model adaptable to post-repair operations. The prognostic model that will be developed is stochastic since stochastic models have already proven their adaptability to unseen test data. However, further development of such models is needed to deal with data on repaired systems. In addition to that, the implementation of a Bayesian Extension allows uncertainty interpretability to be considered to account for the uncertainty coming from the repair action itself but also from the different sources of uncertainties that have not been studied in the field of prognostics.

1. PROBLEM STATEMENT AND STATE-OF-THE-ART

Prognostics and Health Management (PHM) is a field that provides users with a thorough analysis of the health condition of a system which allows users to maximize the operational availability, reduce maintenance costs, and improve the system's reliability and safety (Tsui et al., 2015). PHM includes the following modules: data acquisition, diagnosis, prognosis, and decision-making (Moradi & Groth, 2020). Prognosis takes the information of the data coming from data acquisition alone or both the information of diagnosis and data acquisition. The output of prognosis is then the prediction of the Remaining Useful Life (RUL) of the system, which is the time left before the system reaches failure.

Prognostics plays a vital role in decision-making processes, guiding actions like system retirement or maintenance scheduling. Maintenance strategies vary from perfect maintenance (replacement) to imperfect maintenance (repair), with the latter being favored for its cost-effectiveness (Do Van et al., 2013). (Bougacha et al., 2020) conducted a review on post-prognostic decision-making, particularly focusing on aerospace applications. Existing approaches in this review typically consider current degradation levels or use prognostics assuming the system is as good as new to inform maintenance decisions. (Nguyen & Medjaher, 2019) developed a Deep Learning-based framework that covers the entire process from data-driven prognostics to maintenance decisions. However, the framework's limitation lies in its consideration of only perfect maintenance. To the best of the author's knowledge, (Welz et al., 2017) is the only work that has addressed repair actions in prognostics, emphasizing the importance of including data from repaired systems to enhance prediction accuracy. Yet, this study lacks reporting on RUL prediction and corresponding confidence intervals, providing only an average error of failure time.

Therefore, a significant research gap exists in the current literature regarding how to perform prognostic when the engineering system has been subjected to imperfect maintenance. In other words, there is a need to develop prognostic models that perform accurately when trained on data from systems with no repair but tested on systems repaired one or more times. This gap is notable given PHM's predictive maintenance and cost reduction goal.

Understanding the effects of repair actions on prognostic

models is crucial, as repairs can alter the degradation process of a system. As a consequence, it will negatively affect the performance of the prognostic model by the decrease in the accuracy of RUL predictions, and an increase in the uncertainty of the predictions. Thus, it will reduce the reliability and robustness of prognostics, which will raise concerns about the eligibility of prognostics for decision-making. Therefore, many questions arise to deal with such a scenario. Should prognostic models consider dependencies between pre and post-repair operations? How can the prognostic model acknowledge the health recovery of the system? And how can uncertainty arising from repair actions be managed effectively?

To address the consideration of imperfect repairs in prognostics, it is necessary to develop a robust prognostic model that allows for interpretable uncertainty given the increased uncertainty expected from the repair actions. Understanding the concepts of robustness and uncertainty management, along with the challenges they present, is essential.

Robustness, defined as a system's ability to perform acceptably across various conditions, poses a challenge in prognostics due to the lack of adaptation mechanisms in existing models. Attempts have been made to improve robustness, such as using adaptive batch normalization or domain adversarial neural networks. Still, challenges persist, with high errors in terms of accuracy, along with instability and noise in the predictions

Exceptionally, the Adaptive Non-Homogeneous Hidden Semi-Markov Model (ANHHSMM) demonstrated adaptable capabilities (Eleftheroglou et al., 2020). This stochastic model was trained with 8 composite specimens under fatigue loading, and later on, tested with 3 specimens, also under fatigue loading, and suddenly experienced an unexpected phenomenon. The model provided good results, however, it has not been validated for a case study involving repairs.

Uncertainty management is the second challenge when performing prognostics with data from repaired specimens in the test set. Uncertainty management is defined as the identification of sources of uncertainty and the reduction of uncertainty by leveraging data to better characterize the inherent prognostic uncertainties, thereby reducing their impact on RUL predictions (Sankararaman, 2015).

However, to identify uncertainty it is first necessary to quantify it. Uncertainty quantification (UQ) is already a challenge in data-driven prognostics when using ML models that are deterministic by nature. Such models usually do not report UQ in their RUL predictions, as seen in (Zhu et al., 2020; Ma & Mao, 2020; Ren et al., 2020; Zhang et al., 2023; Cheng et al., 2022). In contrast, some publications address uncertainty quantification when using stochastic models or particle filters, but they provide broad confidence intervals, which results in a lack of valuable information for decision-making (Huang et al., 2017; Cadini, Sbarufatti, Cancelliere, & Giglio, 2019; Cadini, Sbarufatti, Corbetta, et al., 2019; Moghaddass & Zuo, 2014; Liu et al., 2018).

To handle broad ranges of confidence intervals is then necessary to perform uncertainty management. But even though some data-driven prognostic models allow UQ, then it is necessary to identify the sources of uncertainty. The classical categorization divides uncertainty into aleatory and epistemic (Der Kiureghian & Ditlevsen, 2009). However, as the authors themselves have mentioned, such categorization is artificial and it depends mostly on the modeler's choice and the application, which is why it is common to see disagreement on how to disentangle uncertainty by using this categorization.

In (Eleftheroglou et al., 2020), a more relevant categorization for prognostics is proposed, identifying five sources of uncertainty: past uncertainties from manufacturing processes, present uncertainty about the system's health, future uncertainty, model uncertainty, and prediction method uncertainty. This new framework has not been applied to real-life scenarios yet, with existing literature still relying on the classical categorization.

2. Expected Contributions

There is no relevant literature addressing imperfect repair actions from the perspective of prognostics. Therefore, this research will serve as a first attempt to address this issue by developing a robust prognostic model that can be trained with degradation histories of systems that have not been repaired and then tested on degradation histories of repaired systems. Thus, the contribution to the field is a prognostic model that has an adaptation mechanism and can take into account the dependencies between pre and post-repair operation, as well as include the recovery of the system after repair.

Additionally, a Bayesian extension is considered because it allows the estimation of a subjective probability. Unlike the frequentist approach, where the statistics are calculated based on the entire population. This is undesirable since calculating the uncertainty based on the statistics of the entire population when they have been subjected to different conditions has no purpose. Instead, the Bayesian approach works under prior knowledge and available data (Bayarri & Berger, 2004). Even more, the model should include the uncertainty coming from the repair. Identifying this and calculating this source of uncertainty allows for more interpretability in UQ that allows future uncertainty management to have more valuable information for the decision-making process. As mentioned earlier, the classical categorization of uncertainty is not suitable for prognostics. Thus, this research attempts to tackle uncertainty quantification from another perspective that has not been implemented in the literature to date.

3. RESEARCH PLAN

The research plan is divided into three main parts:

- **Experimental Campaign**: since there is no available dataset for prognostics that includes maintenance actions, the first step is to perform an experimental campaign. The experiments consider materials mostly used in aviation structures, such as metals and composites. This phase of the research also involves the analysis of the experimental data, in terms of the effects on the degradation process and a comparison study on how the performance of different prognostic models that are commonly used are affected when dealing with this data.
- **Development of the prognostic model**: As mentioned in Section 1, it is necessary to develop a prognostic model that has an adaptive mechanism. After a literature review, the most suitable model for this application is the ANHHSMM, however, the model needs the addition of variables to take into account the repair of the system as well as the relaxation of some assumptions. Therefore, in this part of the research, the work would consist of developing the mathematical model, including the programming implementation.
- **Bayesian Extension**: Finally, the last part of the research involves the Bayesian extension that allows more interpretable uncertainty in the prognostic model by identifying sources of uncertainty.

As of now, the work that has already been done corresponds to the experimental campaign. The research group performed experiments with open-hole aluminum specimens of material 7075-T6. Each specimen had dimensions $300x45x2$ [mm] and a central hole of 6 [mm] diameter. The aluminum specimens were subjected to constant amplitude fatigue, with a maximum stress of 100 [MPa], frequency of 5 [Hz], and ratio of 0.1. The training data consists of 5 degradation histories of specimens from run to failure. The testing data consists of 5 specimens, also from run to failure. However, the testing specimens were repaired at cycle 14000 with a composite patch to cover the fatigue crack.

Figure 1 shows health indicators derived from experimental data using a neural network developed by the research team. Training trajectories are depicted in blue shades while testing trajectories are in red shades. For visualization, only two trajectories per training and testing set are shown. Notably, a distinct shift in cluster values occurs around cycle 14000 in the testing trajectories, indicating specimen health recovery post-repair. From the plot, it is evident that testing specimens had a longer lifetime, in comparison with the training specimens, due to the repair.

By using this data, a preliminary comparison between prognostic models has been done by the use of SVR and MLP. The results show the poor performance of both of these models

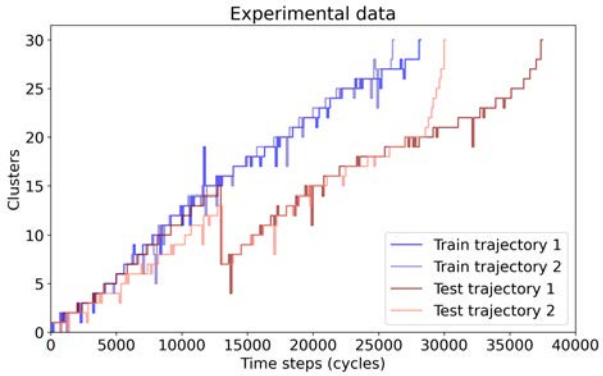


Figure 1. Experimental data of metal specimens for training and testing set.

with an average RMSE value for the test dataset of 131.0119 and 131.4693, respectively. This preliminary comparison shows the lack of adaptability of the models. Future work involves the comparison of more complex prognostic models such as Long Short-Term Memory (LSTM) and the ANHHSMM.

Part of the work in progress, is a literature review on uncertainty quantification in various prognostic models highlights the challenge in data-driven prognostics, particularly with ML models. Despite their high accuracy, ML models struggle with uncertainty quantification due to their deterministic nature. Another limitation is their reliance on the classical categorization of uncertainty into aleatory and epistemic types. The review compares methods for quantifying these uncertainties in ML models and implements a new prognostic measurement for Hidden Markov Models (HMMs) to assess stochastic models' ability to capture relevant uncertainties in prognostics, including past and future sources.

4. CONCLUSIONS

PHM is a field that assesses the health of an engineering system to perform predictive maintenance. Therefore, prognostics are key when predicting the health of the system and give valuable information for decision-making. However, within the prognostic field, a research gap exists when considering maintenance actions, such as repair. Repair is a common procedure that can have an impact on the degradation process of the system, and, therefore, it will negatively impact the performance of a prognostic model if this data is not part of the training set.

This research attempts to develop a robust prognostic model that can be trained with systems that have never been repaired and tested with systems that have been repaired one or several times. Even more, the research will also address UQ challenges such as the quantification of sources of uncertainty under the new categorization allowing more interpretability of uncertainty.

REFERENCES

Bayarri, M. J., & Berger, J. O. (2004). The interplay of bayesian and frequentist analysis.

Bougacha, O., Varnier, C., & Zerhouni, N. (2020). A review of post-prognostics decision-making in prognostics and health management. *International Journal of Prognostics and Health Management*, *11*(15), 31.

Cadini, F., Sbarufatti, C., Cancelliere, F., & Giglio, M. (2019). State-of-life prognosis and diagnosis of lithium-ion batteries by data-driven particle filters. *Applied energy*, *235*, 661–672.

Cadini, F., Sbarufatti, C., Corbetta, M., Cancelliere, F., & Giglio, M. (2019). Particle filtering-based adaptive training of neural networks for real-time structural damage diagnosis and prognosis. *Structural Control and Health Monitoring*, *26*(12), e2451.

Cheng, Y., Hu, K., Wu, J., Zhu, H., & Lee, C. K. (2022). A deep learning-based two-stage prognostic approach for remaining useful life of rolling bearing. *Applied Intelligence*, *52*(5), 5880–5895.

Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, *31*(2), 105–112.

Do Van, P., Voisin, A., Levrat, E., & Iung, B. (2013). Remaining useful life based maintenance decision making for deteriorating systems with both perfect and imperfect maintenance actions. In *2013 ieee conference on prognostics and health management (phm)* (pp. 1–9).

Eleftheroglou, N., Zarouchas, D., & Benedictus, R. (2020). An adaptive probabilistic data-driven methodology for prognosis of the fatigue life of composite structures. *Composite Structures*, *245*, 112386.

Huang, Z., Xu, Z., Ke, X., Wang, W., & Sun, Y. (2017). Remaining useful life prediction for an adaptive skew-wiener process model. *Mechanical Systems and Signal Processing*, *87*, 294–306.

Liu, T., Zhu, K., & Zeng, L. (2018). Diagnosis and prognosis of degradation process via hidden semi-markov model. *IEEE/ASME Transactions on Mechatronics*, *23*(3), 1456–1466.

Ma, M., & Mao, Z. (2020). Deep-convolution-based lstm network for remaining useful life prediction. *IEEE Transactions on Industrial Informatics*, *17*(3), 1658–1667.

Moghaddass, R., & Zuo, M. J. (2014). An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process. *Reliability Engineering & System Safety*, *124*, 92–104.

Moradi, R., & Groth, K. M. (2020). Modernizing risk assessment: A systematic integration of pra and phm techniques. *Reliability Engineering & System Safety*, *204*, 107194.

Nguyen, K. T., & Medjaher, K. (2019). A new dynamic predictive maintenance framework using deep learning for failure prognostics. *Reliability Engineering & System Safety*, *188*, 251–262.

Ren, L., Dong, J., Wang, X., Meng, Z., Zhao, L., & Deen, M. J. (2020). A data-driven auto-cnn-lstm prediction model for lithium-ion battery remaining useful life. *IEEE Transactions on Industrial Informatics*, *17*(5), 3478–3487.

Sankararaman, S. (2015). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing*, *52*, 228–247.

Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., Wang, W., et al. (2015). Prognostics and health management: A review on data driven approaches. *Mathematical Problems in Engineering*, *2015*.

Welz, Z., Coble, J., Upadhyaya, B., & Hines, W. (2017). Maintenance-based prognostics of nuclear plant equipment for long-term operation. *Nuclear Engineering and Technology*, *49*(5), 914–919.

Zhang, X., Sun, J., Wang, J., Jin, Y., Wang, L., & Liu, Z. (2023). Paoltransformer: Pruning-adaptive optimal lightweight transformer model for aero-engine remaining useful life prediction. *Reliability Engineering & System Safety*, *240*, 109605.

Zhu, J., Chen, N., & Shen, C. (2020). A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. *Mechanical Systems and Signal Processing*, *139*, 106602.

Trustworthy Machine Learning Operations for Predictive Maintenance Solutions

Kiavash Fathi[1,2], Tobias Kleinert[2], Hans Wernher van de Venn[1]

[1] *Institute of Mechatronic Systems, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland*
fath@zhaw.ch, vhns@zhaw.ch

[2] *Chair of Information and Automation Systems for Process and Material Technology, RWTH Aachen University, 52064 Aachen, Germany*
kiavash.fathi@rwth-aachen.de, kleinert@plt.rwth-aachen.de

ABSTRACT

With the ever-growing capabilities of data acquisition and computational units in industry, development, and deployment of data-driven models (*e.g.*, predictive maintenance solutions) have become more abundant. However, if these models are not trained and maintained properly, they can be counterproductive as their predictions may be incorrect, unreliable, or difficult to interpret. In addition, unlike conventional software, the issues with such models often result in reduced productivity rather than traceable software errors. Therefore, we aim to use model performance evaluation measures introduced in trustworthy AI operations (TrustAIOps) to trigger re-evaluation of different parts of the data pipeline and the deployed data-driven model given machine learning operations (MLOps) requirements. We argue that by creating an ecosystem capable of monitoring different aspects of a data-driven solution by integrating and managing the implementation concepts in TrustAIOps and MLOps, it is possible to boost the performance of models given the constant changes induced by the specifications of Industry 4.0.

1. INTRODUCTION

Data acquisition and computational units improve daily which facilitate the development and deployment of data-driven approaches in Industry 4.0 settings. However, these data-driven models, when not trained and maintained properly, can be counterproductive as their predictions are not correct, reliable or interpretable. Unlike conventional software, the issues with model development manifest themselves in reduced productivity and not in other forms of traceable software error. In fact, when faced with during the run-time, they could be due to the errors from the data acquisition, data preprocessing, model training or model deployment submodels (Ashmore, Calinescu, & Paterson, 2021).

To ensure the acceptable performance of data-driven solutions, numerous implementation concepts have been introduced from the machine leaning operations (MLOps) society, which cover different aspects of preparing and deploying a data-driven solution. The following are some the most important characteristics of the models developed given MLOps requirements (Huyen, 2022):

1. **Reliability:** Correctness despite adversity
2. **Scalability:** Possibility of growth in complexity
3. **Adaptability:** Can cope with different data distribution shifts and business requirements
4. **Maintainability:** Documented and open to different tools

On the other hand, given the ever-growing application of machine learning solutions in different use cases, especially in safety-critical systems, performance criteria other than the accuracy have been promoted in research targeting trust worthy AI operations (TrustAIOps) which include but are not limited to (Li et al., 2023):

1. **Robustness:** Ability to deal with unseen data
2. **Generalization:** Distilling knowledge from limited training data for accurate predictions on unseen data
3. **Explainability:** Clarity on how a model makes decision
4. **Transparency:** Disclosing information about the model's lifecycle

As it can be seen in the above-mentioned characteristics, both MLOps and TrustAIOps put much emphasis on the performance of the deployed models given the possible changes in the data. These changes in Industry 4.0 settings are also relevant as there are many factors including production recipe, raw material vendor, product test unit fail/pass criteria, asset wear and tear, *etc.*, which can cause different types of data

distribution shifts. Predictive maintenance (PdM) as one of the important use cases of Industry 4.0 compliant solutions, is not an exception and requires tailored solutions for ensuring its effectiveness in an industrial setting.

1.1. Problem statement

How can the model performance evaluation measures introduced in TrustAIOps be used to trigger re-evaluation of different parts of the data pipeline and the deployed model given MLOps requirements? *(As a small remark; however, given the fact that the MLOps and TrustAIOps requirements cover numerous aspects of the PdM models, in the conducted studies, we consider only the characteristics listed above.)*

In the conducted research, we aim to introduce new implementation concepts which have proven to be useful for real industrial use cases in Europe and that are not properly addressed in the related work. In what follows pairs of MLOps and TrustAIOps, written as

$$TrustAIOps\ trigger \rightarrow MLOps\ requirement$$

are introduced with a specific implementation challenge for industrial PdM solutions:

1. **Robustness → Reliability:** Detecting previously unseen failures in the system
2. **Explainability → Scalability:** Interpretable model stacking
3. **Generalization → Adaptability:** Classifying different working conditions of an asset - Generation of run-to-failure data via simulation models
4. **Transparency → Maintainability:** Human-readable reports from different parts of the PdM solution

1.2. Research questions (RQs) and expected contributions

To elucidate further, given the complexity and high dimensionality of industrial data from different assets, how can

RQ 1. The model prediction certainty be correctly interpreted for out-of-training-distribution datapoints which represent previously unidentified failures of an asset? (see red blocks in Fig. 1). For inspecting the data-distribution shifts caused by changes in the working conditions refer to **RQ 3**.

RQ 2. The impact of different sources of uncertainty be minimized during model training using interpretable AI? (see grey blocks in Fig. 1)

RQ 3. Domain knowledge about different working conditions be included in data preprocessing and model training for enhanced data aggregation across different instances of the same production assets? (see green blocks in Fig. 1)

RQ 4. Lack of annotated data, *e.g.*, continuous data such as run-to-failure samples, be compensated using domain adaptation and simulation models? (see blue blocks in Fig. 1)

RQ 5. Human-readable reports be generated for increasing the transparency, *e.g.*, about how predictions are made and what data was used to train the model, of different submodels of the PdM solution, *esp.* for safety-critical system?

2. Conducted studies

In this section, a summary of the implemented solutions targeting parts of the first four **RQs**, specifically developed for the industry are presented. The solutions provided in this section adhere to the identical sequence as outlined in the **RQs**.

2.1. Detecting previously unidentified failures of an asset (Industry supported academic project)

It has been shown that the available data from different assets, even in case that they are abundant, normally do not cover different failure types that could occur in a system. Therefore, it is inevitable to monitor a PdM model in case data from a new working condition and/or failure type are exposed to it (Fig. 2). Despite numerous model calibration solutions, it has been observed that even models which are calibrated cannot demonstrate their certainty correctly when out-of-training-distribution data are fed into them. For PdM solutions, it is of utmost importance to inform the maintenance crew when a novel in the system has occurred as, otherwise, an exhaustive search is require for fault localization and diagnostics. In the conducted study, we have developed a post-hoc sample-based classification model built on top of the initial PdM solution that can detect previously unidentified failures in the system. The proposed method inspect the behavior of the PdM model, defined as the sequence of the PdM model certainty, and flags datapoints which indicate an anomaly in the PdM model behavior. The proposed method is tested on a demonstrator build by a company producing pneumatic components and has a mean accuracy of 94.35% (Fathi, Ristin, Sadurski, Kleinert, & van de Venn, 2024).

2.2. Reducing model uncertainty by interpretable model stacking (Industrial project)

Various changes in the production, *e.g.*, recipe updates, raw material vendor changes, improvements in quality test unit fail/pass criteria, *etc.*, impact the performance of the trained models given the potential data distribution shifts. If fact, with the adaptability in production as one of the main focuses of Industry 4.0, these changes reflect themselves in the data gathered from different assets which directly can impact the quality of the production. It is possible to counteract these changes in the gathered data by using different ensembling and model stacking techniques. In the conducted study we propose a novel approach for stacking the formerly trained

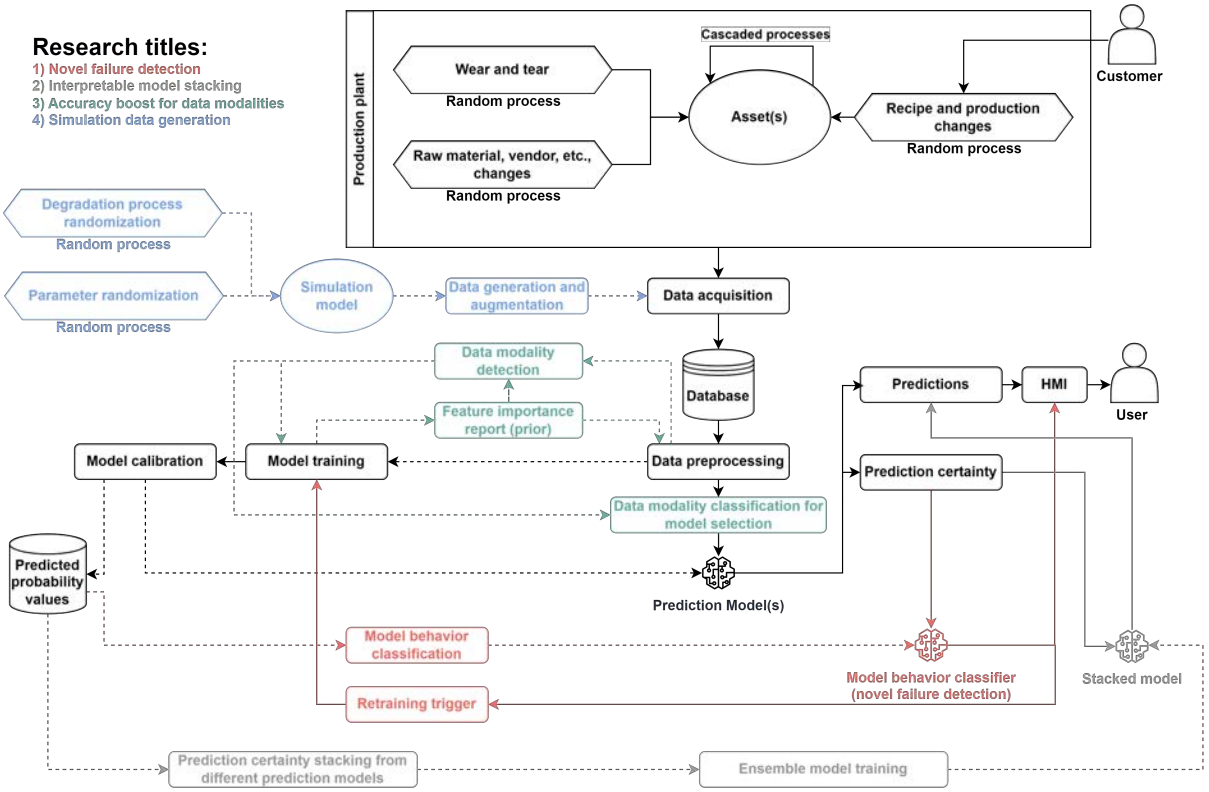


Figure 1. Overview of the proposed solution for TrustAIOps and MLOps integration in PdM

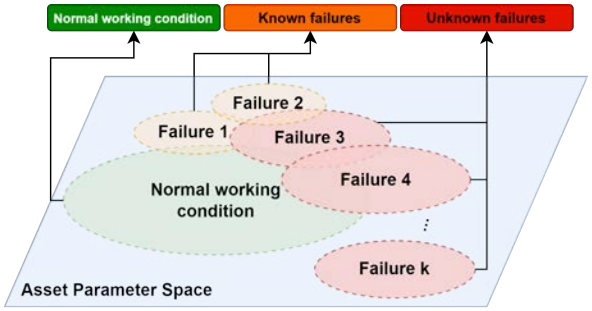


Figure 2. Asset parameter space and different known and unknown data modalities of the system

base learners. To avoid information loss due to prediction quantization of the base learners, in the proposed method we directly use the predicted probability values from the base learners and stack them using a linear regression model. The results demonstrate a 19.49% reduction in the binary estimated calibration error compared to convectional models which indicates the increased reliability of the final solution (Fathi, Stramaglia, et al., 2024).

2.3. Boosting model accuracy for different data modalities of an asset (Industrial project)

The constant changes in the production introduced in Subsection 2.2, can also lead to different dominant working condition of an asset which is also refereed to as data modality. In the conducted study, two instances of the same milling machine used for creating artificial bone joints of different sizes are examined to first detect and later to classify their different data modalities (see Fig. 3). Once different data modalities are distinguishable from one another, separate prediction models are trained for them which can increase the overall accuracy of the predictions up to 25.20%. In addition, for the data modality which forms the minority of the data from the asset, it is shown that by combining the corresponding data modalities from the two milling machines, it is possible to increase the accuracy for the aforementioned data modality up to 60.50%. In fact, by detecting corresponding data modalities, it is possible to address the problem of lack of annotated data for different instances of the same asset by simply sharing data from the same data modalities across the assets (Fathi, Sadurski, Kleinert, & van de Venn, 2023).

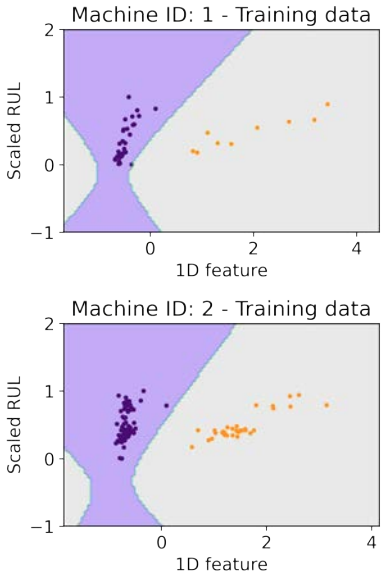


Figure 3. Decision boundaries of the trained model for classifying different data modalities of the asset

2.4. Data generation from simulation model for domain adaptation (Industrial project, paper under review)

Domain adaptation techniques developed for PdM normally focus on classification problem and neglect the regression problem of estimating the remaining useful life of an asset. In addition, they do not consider cases where the degradation of the asset is a random process itself either given the possibility of changes in the dominant failing component. Therefore, in the conducted study a novel approach for simulation data generation is introduced which is based on simulation parameter and data perturbation. It is shown how the proposed method can help cover different regions of the parameter space of the asset indicating different working conditions and parameterization of the asset (see Fig. 4). As a result, models trained with such data are more robust against signal reading manipulation and also demonstrate a more spread-out feature importance across a wider range of sensor readings while making predictions.

3. Future work and next steps

Given the conducted studies listed above, it is inevitable to create an ecosystem which is capable of monitoring different aspects of a PdM solution by **integrating** and **managing** the implementation concepts introduced in Section 2. In fact, this ecosystem will use the introduced TrustAIOps concepts to ensure the expected performance of the PdM solution given MLOps requirements. One of the most important features of this ecosystem as introduced in **RQ 5** (see Section 1.2), is providing human-readable reports from different submodels of the PdM solution to ease its maintenance and debugging. One feasible solution for the aforementioned ecosystem is to create a metadata-based management system which is capable of tracking changes in different submodels of the deployed PdM solution. These changes are the essentially the response of the PdM solution for adapting to the new working conditions and/or previously unseen failures of the system. When done correctly, the proposed solution can be used as a foundation for data-driven PdM solutions of different assets including safety-critical systems.

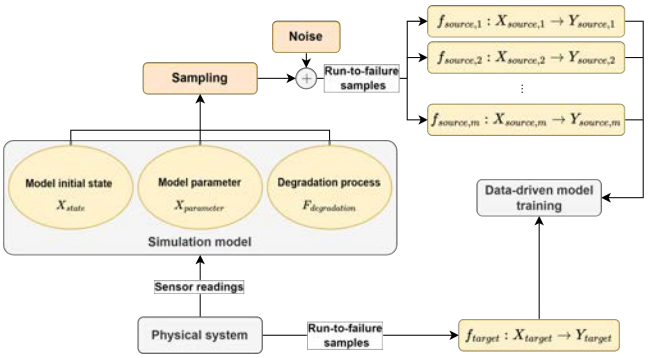


Figure 4. Domain adaptation via simulation parameter and data perturbation

References

Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, *54*(5), 1–39.

Fathi, K., Ristin, M., Sadurski, M., Kleinert, T., & van de Venn, H. W. (2024). Detection of novel asset failures in predictive maintenance using classifier certainty. *IEEE, 32nd Mediterranean Conference on Control and Automation (MED)*.

Fathi, K., Sadurski, M., Kleinert, T., & van de Venn, H. W. (2023). Source component shift detection classification for improved remaining useful life estimation in alarm-based predictive maintenance. *IEEE, 23rd International Conference on Control, Automation and Systems (ICCAS)*.

Fathi, K., Stramaglia, M., Ristin, M., Sadurski, M., Kleinert, T., Schönfelder, R., & van de Venn, H. W. (2024). Sustainability in semiconductor production via interpretable and reliable predictions. *IFAC, 12th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*.

Huyen, C. (2022). *Designing machine learning systems*. "O'Reilly Media, Inc.".

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46.