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.

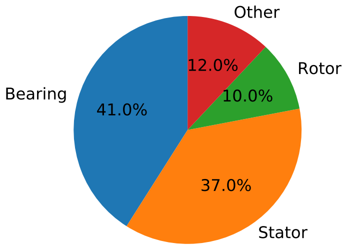

Figure 1. Proportion of machine components failure

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, aeroengines, 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

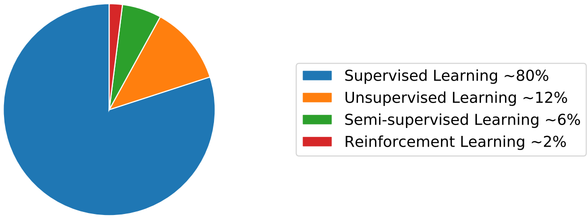

Figure 2. ML techniques used for MFD

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 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.

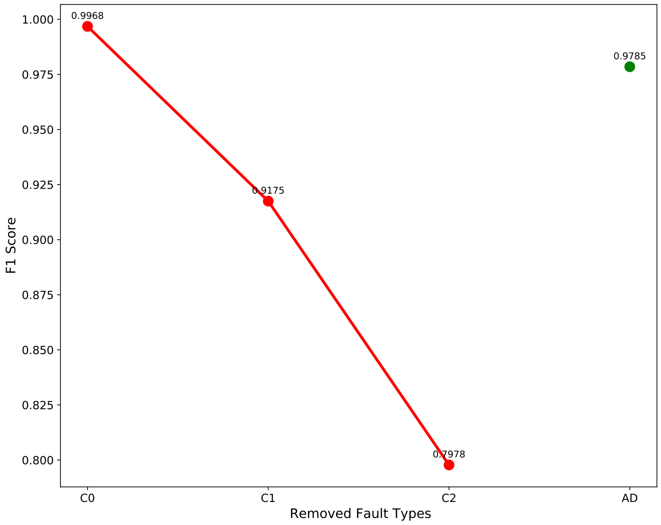

Figure 3. Average F1 score of the ResNet-DT classifier from scalogram images of vibration signals on the CWRU dataset.

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 9̃8% 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.

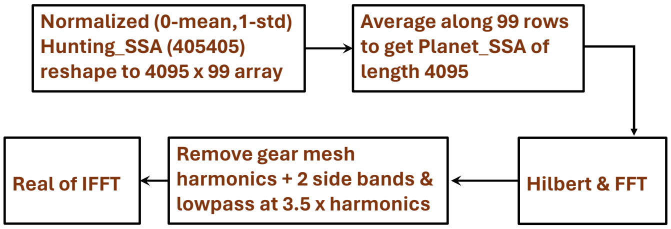
Normalized (0-mean,1-std) Hunting_SSA (405405) reshape to 4095 x 99 array
Average along 99 rows to get Planet_SSA of length 4095
Hilbert & FFT
Remove gear mesh harmonics + 2 side bands & lowpass at 3.5 x harmonics
Real of IFFT

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	$\mu - 3\sigma$

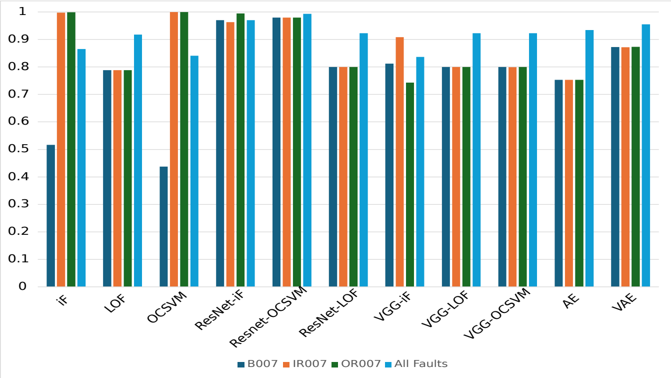

Figure 5. Comparison of Models' Performances on the CWRU Dataset.

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 committee³, 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.

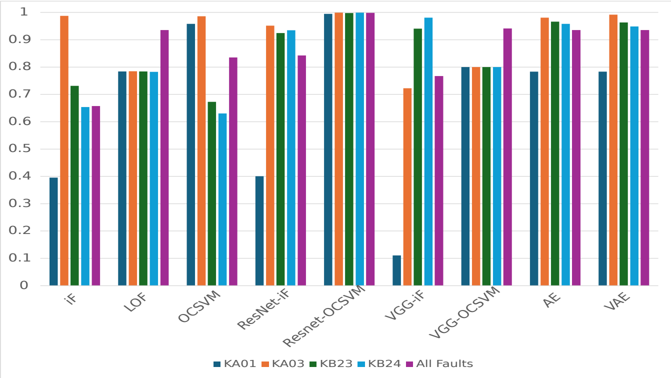

Figure 6. Comparison of Models' Performances on PU Dataset

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

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

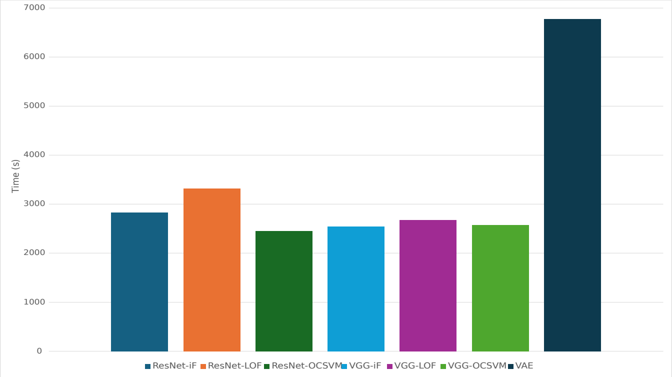

Figure 7. Comparison of Model Performances Based on Runtime: The figure illustrates the time taken by various models to complete 10 runs of anomaly detection using all 4 test sets from the CWRU dataset. Notably, all models operate on the same input, namely, scalograms.

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