PIAFC: Physics-Inspired Anomaly-based Fault Classification

Amirhossein Berenji, Sławomir Nowaczyk, Zahra Taghiyarrenani

[a]*Center for Applied Intelligence Systems Research (CAISR), Halmstad University, Kristian IV:s väg 3, Halmstad, Sweden*

Abstract

Intelligent fault diagnosis (IFD) solutions have garnered significant attention in academia; however, limited access to faulty data and the black-box nature of most effective models hinder their practical implementation in industries. A straightforward solution to the challenge of black-box models is to employ eXplainable Artificial Intelligence (XAI) methods, which can provide insights into their inner workings. This paper aims to showcase an additional area where XAI methods can be beneficial. We introduce an XAI-powered fault classification approach that not only detects anomalous operation but also matches it with the most probable failure mode. The key advantages of employing our proposed pipeline are that it can be implemented using fault-free datasets, provides high transparency due to human-understandable decision-making strategies, and has multi-component fault classification capability. We showcase the efficacy of our method, called Physics-Inspired Anomaly-based Fault Classification, using bearing fault classification use cases, including transfer learning scenarios. However, it is generalizable to all fault diagnosis tasks that exhibit physically defined fault patterns.

Keywords: Anomaly Detection, Fault Classification, Autoencoder, Rolling Element Bearing, Vibrations, eXplainable AI, Transfer Learning

1. Acknowledgments

This work was partially supported by the Knowledge Foundation, Vinnova (Sweden's innovation agency) through the Vehicle Strategic Research and Innovation Programme FFI.

PIAFC: Physics-Inspired Anomaly-based Fault Classification

Abstract

Intelligent fault diagnosis (IFD) solutions have garnered significant attention in academia; however, limited access to faulty data and the black-box nature of most effective models hinder their practical implementation in industries. A straightforward solution to the challenge of black-box models is to employ eXplainable Artificial Intelligence (XAI) methods, which can provide insights into their inner workings. This paper aims to showcase an additional area where XAI methods can be beneficial. We introduce an XAI-powered fault classification approach that not only detects anomalous operation but also matches it with the most probable failure mode. The key advantages of employing our proposed pipeline are that it can be implemented using fault-free datasets, provides high transparency due to human-understandable decision-making strategies, and has multi-component fault classification capability. We showcase the efficacy of our method, called Physics-Inspired Anomaly-based Fault Classification, using bearing fault classification use cases, including transfer learning scenarios. However, it is generalizable to all fault diagnosis tasks that exhibit physically defined fault patterns.

Keywords: Anomaly Detection, Fault Classification, Autoencoder, Rolling Element Bearing, Vibrations, eXplainable AI, Transfer Learning

1. Introduction

The need for higher reliability, safety, and performance, alongside more affordable sensors – for measurement – and network infrastructures – for data transportation [1], has highlighted condition-based maintenance as a promising maintenance strategy for crucial industrial machines. Among all, rotating machines are one of the most critical equipment types [2]; thus, a lot of attention is paid to extending their uptime and reliability through the application of intelligent methods in condition monitoring, known as

Intelligent Fault Diagnosis (IFD) and defined as *"the applications of machine learning theories to machine fault diagnosis"* [3].

Although IFD has been successful through condition-based maintenance of certain equipment, its application is firmly constrained to scenarios where vast amounts of data describing both faulty and normal operation are available. Contrarily, extensive data collection from industrial machines during faulty operation is impossible in most real-world applications, not only because faults are supposed to occur rarely but also because operating industrial machines in a defective state is not allowed. Thus, developing IFD solutions implementable with limited faulty data is one of the main trends in IFD.

Various approaches are introduced to cope with the limited faulty data challenge, from contrastive representation learning to meta-learning and transfer learning. While the typical motivations behind incorporating physical knowledge in IFD applications are increasing interpretability [4, 5] and performance improvement [6], we also take it as a source capable of compensating for the faulty data unavailability. Traditionally, fault diagnosis of these machines includes searching the vibration frequency spectrum for fault characteristic frequency components, i.e., frequency components whose appearance corresponds to the occurrence of specific faults. Therefore, faults are recognizable according to the frequency ranges with dominant peaks, which is a rich source of information.

In this paper, we aim to use physical knowledge – in the form of bearing specifications and shaft rotational speed – to take on the limited faulty data challenge. In particular, we introduce a method to pseudo-label faulty frequency-domain vibration signals without using any faulty data. Our method first identifies anomalies, i.e., data that differs significantly from regular operation. Then, it compares each faulty signal against the potential fault signatures and identifies the most similar fault pattern. The main advantages of this method, are: 1) **Training-free**: no training is required, therefore no training data is needed and 2) **Inherent interpretability**: it is powered by physical knowledge and the similarity of the given signal with the expected fault patterns, so it is naturally interpretable. Next, we employ eXplainable Artificial Intelligence (XAI) to highlight frequency bands that contribute most towards anomalous predictions. Comparing these frequency bands with the physically expected fault characteristic components, a human practitioner will be assisted in distinguishing the fault type, again using no faulty data.

We demonstrate the efficiency of this pipeline using a bearing fault clas-

sification setup, as they are arguably the most critical component within rotary machines. Their importance can be better understood by figures; around 45 to 55 % of the failures in rotary machines are due to bearing problems [7]. When it comes to bearings, for each fault type of the bearing – e.g. inner/outer race fault, ball problem, and cage problem – a ratio is defined according to its geometry; once a fault comes into being, dominant peaks appear within the vibration frequency spectrum at the corresponding fault characteristic frequencies – which is the multiplication of the shaft running speed and the fault-specific ratios. It is noteworthy that although our case study is conducted on bearing fault classification problems, it is generalizable to any set of sufficiently distinguishable faults of rotating machines, through vibration analysis.

The main contribution of this paper is introducing a method to detect and classify bearing faults, with no faulty data required. The proposed method fuses component-level physical knowledge and XAI-generated insights of data-driven models, to assist human practitioners in the fault diagnosis of rotary machines. Secondly, we showcase its advantage in multi-component setups, rather than conventional single-component scenarios. Last but not least, we also evaluate its potential to be used in a transfer learning playground, where the source and target components are different.

The rest of this paper is organized as follows: in Section 2, we review the most relevant studies in the literature. Next comes the description of the proposed method, in Section 3. Following that, in Section 4 we showcase the capabilities of the proposed method in both single-component and multi-component scenarios. Throughout the section 5, we investigate different aspects of the proposed method by conducting multiple ablation studies. Last but not least, in Section 6 we discuss our findings comprehensively and conclude.

2. Related Works

To better organize this section, we divide it into four parts; we start by investigating previous works on bearing fault classification with limited data in Section 2.1. Next, in 2.2, we present the most similar studies to ours from the perspective of anomaly detection. Moreover, in section 2.3 we introduce the previous trials to take advantage of physical understanding for bearing IFD purposes. Finally, in 2.4 we provide examples of studies where XAI methods are employed to explain black-box IFD solutions.

2.1. Limited-data Bearing Fault Classification

As mentioned earlier, bearings are among the most critical components of rotary machines; thus, huge attention is paid to their fault classification. However, the main limitation to employing intelligent methods to diagnose them is the requirement of huge datasets. To overcome this, researchers try to develop bearing fault classification solutions, that can handle limited-data scenarios. Meta-Learning is one approach to tackle this problem; an example is [8] where authors employ model-agnostic meta-learning (MAML) [9] for few-shot bearing fault classification. They not only demonstrate that MAML outperforms the previous state-of-the-art (contrastive representation learning by Siamese Networks) significantly, but also show that it is capable of recognizing real bearing faults, using the knowledge acquired from artificially created faults. Similarly, authors of [10] also use MAML to achieve high accuracy and fast convergence in few-shot bearing fault classification within complex working conditions. They first transform raw time signals into time-frequency representations using Short-Time Fourier Transform (STFT); next, they use the convention introduced in [9] to build a variety of classification tasks, with different working conditions. During the meta-training process, their model gains prior knowledge on optimizing the model weights by fitting training tasks. During the meta-testing process, the model learns how to fit tasks from unseen working conditions. According to experiments, their proposed method outperforms a wide range of classifiers; from k-Nearest Neighbor to ResNet and matching networks. Alternatively to those, in [11] authors use a metric-based Meta-Learning approach, known as Reinforce Relation Network (RRN), to diagnose bearing faults in limited training data scenarios. To cope with the overfitting vulnerability in limited data problems, authors use label smoothing; additionally, they use the Adabound optimizer to adjust the learning rate, dynamically. They evaluate the performance of their proposed method with Support Vector Machines (SVMs) and Transfer Learning scenarios and prove the superiority of their method. Yet another approach is to use contrastive representation learning; for instance, in [12], authors employ Simames Networks with wide kernel convolutional neural networks (WDCNN) as the backbones to extract highly discriminative feature sets from acceleration signals to diagnose bearings. They compare their method with multiple conventional approaches, including SVM and WDCNN classifiers with conventional training, demonstrating the superior performance of their approach. Additionally, in [13], authors successfully cope with the few-shot learning problem by transfer-

ring knowledge across different bearings. Their experiments not only show the superiority of their method over various baselines but also successful knowledge transferring in complex transfer scenarios (e.g., different working conditions and cross-machine knowledge transfer). Generally, unlabeled data is considered to be easier (and more affordable) to collect, when compared to labeled faulty data. Thus, [14] combined unsupervised feature learning by autoencoder training to take advantage of the unlabeled data and used contrastive learning by Siamese networks to fine-tune the encoder. That method was compared with different classification approaches (conventional classification, contrastive representation learning and conventional fine-tuning of the encoder), where results demonstrate that the hybrid method improves accuracy. Additionally, the supplementary experiments demonstrate that freezing the initial layers of the encoder during the fine-tuning process can boost the performance, noticeably.

2.2. Anomaly Detection in Rotary Machines

Anomaly detection is arguably the most fundamental task in the condition-based maintenance of rotary machines; as a result, numerous studies are centered around it. As an example, in [15], anomaly detection of the IMS bearing dataset is approached as a supervised classification task, and the performance of SVM and One-Class Support Vector Machines (OCSVM) is compared. Their findings suggest that OCSVM is likely to outperform conventional SVM in binary classification problems. It is worth mentioning that feature extraction is done using statistical features from both time and frequency domain signals and Principal Component Analysis is employed for feature selection. Similarly, in [16], a combinatory approach employing Artificial Neural Network (ANN) and isolation forest is introduced. In their method, the ANN classifier is trained in a supervised manner, first and then, the hidden layer embedding borrowed from the network is used as the input of the isolation forest. The authors demonstrated the effectiveness of their proposed method by comparing its performance with traditional anomaly detection methods in a gearbox vibration analysis case study; where their method outperforms its competitors.

Autoencoders are great candidates to solve anomaly detection tasks and their application within IFD is well studied, due to the following reasons: 1) Deep autoencoders are capable of automatic feature extraction from unstructured data, e.g., raw time signals, frequency spectra, and spectrograms, 2) Anomaly detection autoencoders are applicable to fault-free datasets too,

as they require no anomalous data in the training stage. To name an example, authors in [17] employed an autoencoder with Long short-term memory (LSTM) blocks to detect bearing anomalies using raw time-domain vibration signals. Similarly, in [18] authors employed Continuous Wavelet Transform (CWT) – to transform given signals into images – alongside convolutional autoencoders to detect anomalies from the time-domain representation of acceleration signals. The reconstruction loss between the original signal and its reconstructed version is used as the indicator, discriminating the normal and anomalous behavior of the machinery. Another example of using autoencoders to manipulate mechanical vibrations is [19], where a convolutional deep autoencoder is used to detect anomalous behavior of a conveyor belt setup, according to the spectrograms from vibration signals. Authors replaced conventional autoencoder training with a masked autoencoder training procedure – in which the autoencoder learns to reconstruct a noise-reduced version given a noisy record – to improve the robustness of their model.

Autoencoders also perform well in anomaly detection within the manually-engineered feature spaces; in [20], authors have compared the performance of deep autoencoders with several other anomaly detection approaches, to differentiate between the normal and faulty operation of bearings. The input utilized in their study is a feature set including time domain, frequency domain, and time-frequency domain features of bearing acceleration signal. Similarly, authors of [21] employed autoencoders to detect anomalous behavior of a sliding bearing. The input to the autoencoder is a four-element long column vector including Root Mean Square (RMS), variance, energy and counts of peaks higher than the pre-defined threshold of acoustic emission signals. Once the autoencoder recognizes an anomalous record, its spectrogram is passed to a deep convolutional classifier to identify the fault type. Unlike previously referenced studies where the input is a representation of a single property, in [22] instantaneous multi-point measurements of different sensors – including vibrometers, pressure sensors, and temperature sensors – is used as the input of an autoencoder to detect the anomalous operation of a gas turbine.

2.3. Physics-informed IFD for bearings

Recently, the application of physics-informed machine learning for intelligent maintenance of rotary machines, particularly for bearings, has received noticeable attention. To name an example, in [23], authors introduce

a physics-based convolutional neural network (PCNN); they employ spectral kurtosis and envelope analysis to extract fault-related signatures from raw signals and feed this enriched representation into the CNN model. They compare PCNN with both traditional machine learning and conventional deep learning approaches of IFD and showcase its superiority. Moreover, PCNN offers multiple bearings monitoring concurrently, as it can process multi-sensor information. Similarly, in [4], authors employ a continuous wavelet convolutional layer as the initial layer to make conventionally black-box CNNs interpretable. This way, the kernels learned in this layer will be able to discover physically meaningful relations. Unlike the randomly initialized kernels in conventional convolution layers, the filter bank in this layer consists of parametrized wavelet dictionaries, accomplishing wavelet transform. According to their experiments on different case studies – bearing fault, helical gear fault and aeroengine bevel gear fault – their proposed method outperforms its conventional competitors. Moreover, in [5] authors introduce an innovative feature weighting layer to make the model more sensitive towards frequency ranges closer to the fault characteristic frequencies. This way, the model will be able to overcome the domain shift due to unseen operation conditions, by extracting robust fault-related features. This model is deployed to an edge computing device for a real-time evaluation of the bearing. A physical understanding of degradation mechanisms can also be used alongside pure data-driven approaches. One example is [24], where authors develop a hybrid model consisting of two different layer types: physics-informed layers – responsible for modeling relatively well-understood degradation patterns – and pure data-driven layers – to model physically complex degradation patterns. The proposed method is evaluated on a wind turbine bearing fatigue failure prediction case study, and they show that this method not only incorporates the physical understanding of bearing life degradation pattern but also successfully compensates for the poor understanding of the grease degradation through pure data-driven modeling. Subsequently, in [25] authors replace the accurately collected laboratory lubricant state indexes with a categorical grease condition descriptor assessed through the visual inception of the lubricant by a technician and naked eye; this way, expensive costs of laboratory analysis are avoided.

2.4. Explainability in IFD

The hunger for maximum uptime preservation alongside the unfortunate consequences of unexpected failures, brings up the need for making IFD so-

lutions understandable to humans. Therefore, the application of XAI techniques to explain complex IFD solutions has gained special attention. For example, authors of [26] employed Gradient Class Activation Mapping (Grad-CAM) to explain the predictions made by a bearing fault diagnosis classification network. The referenced network utilizes images derived from the STFT as the input to the network, and the classification network takes advantage of convolutional layers in its initial layers to extract more abstract feature sets. Similarly, in [27], authors use a variant of Shapley Additive Explanation, known as Kernel SHAP, to explain a k-Nearest Neighbor classifier. They compare their proposed method with conventional classification approaches and demonstrate comparative classification performance alongside improved interpretability. Another example is [28], where different XAI methods, including Class Activation Maps (CAM) and saliency maps, are employed to explain a deep convolutional classifier. This classifier is provided with triaxial vibration records describing the operation of a gearbox. According to the results presented by the authors, their proposed method not only outperforms traditional classification methods such as SVM and decision trees according to classification performance metrics but also provides a higher level of interpretability. Additionally, in [29], authors introduce a multi-step procedure capable of machinery fault classification in an unsupervised manner; their approach consists of anomaly detection and in case they detect anomalous behavior they pursue the application of XAI methods to extract feature importance. Collected feature importance scores are then used to match the anomalous data record with the expected fault patterns by the expert and determine fault class; for the cases where the extracted explanation matches can be associated with more than one fault class, authors suggest a root cause analysis to identify the fault class. The effectiveness of their method is evaluated over different rotary machine vibration datasets and compared with different machine learning methods.

The need for domain-specific explanations in IFD resulted in the creation of XAI methods specially designed for this domain. As an example, in [30] authors introduce the Frequency Activation Maps (FAM) method to highlight the frequency band in which a one-dimensional deep convolutional classifier focuses to determine whether a vibration signal is associated with normal or faulty operation. Their experiments show that this method demonstrates the model classification criteria perfectly and fault characteristic frequencies are identifiable.

3. Methods

In this section, we first start by providing an overview of our proposed method; next, we will briefly introduce the different building blocks, it consists of.

3.1. Proposed method

A visual illustration of the proposed method and how different building blocks fit together is presented in Figure 1. Physics-Inspired Anomaly-based Fault Classification (PIAFC) consists of three modules: data preprocessing, anomaly detection, and fault classification. As the name indicates, the data preprocessing module is responsible for the employment of basic data preparation operations, including Hilbert Transform (HT) for demodulation, Zoom FFT to extract high-resolution frequency domain representation, and min/max scaling to scale the frequency domain signal. Following that, the preprocessed record is fed to an autoencoder as the anomaly detection module. Comparing its reconstruction error (Rec_{err}) with the predefined threshold, normal and anomalous signals are distinguishable. For Rec_{err} lower than the threshold, the record is specified as normal, and no further analysis is required; however, for Rec_{err} values higher than the threshold – anomalous ones – the fault classification module determines the fault type.

The fault classification module includes two separate steps; firstly, the anomalous spectrum is compared with the expected fault patterns to determine the most probable fault class, using the similarity-evaluation property of the inner product. Secondly, we explain the anomaly detection autoencoder prediction by XAI to provide a complementary source of information, highlighting the frequency ranges most contributing to the Rec_{err}. These explanations as the complementary information source, not only highlight the most disturbing frequency ranges to assist human practitioners in recognizing the fault type, but also work as a backup mechanism to verify the validity of the identified class by the first step.

It is noteworthy that although the anomaly detection and fault classification modules are both well-understandable to human practitioners; they belong to contrasting schools of thought on XAI. On the one hand, the anomaly detection module is explained using post-hoc explanation methods; on the other hand, the fault classification module takes advantage of an inherently transparent decision-making approach, falling under the class of interpretable machine learning.

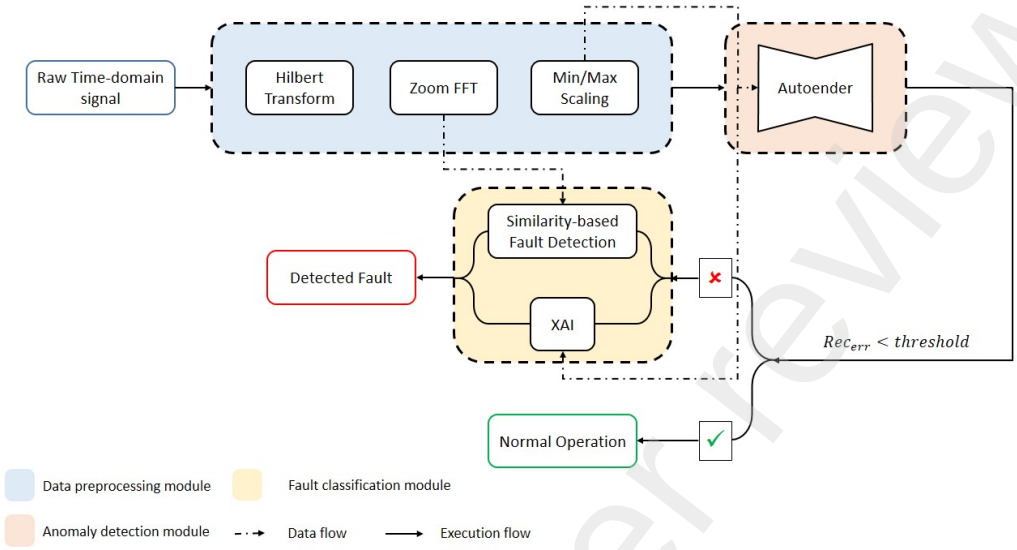


Figure 1: Visual illustration of the proposed method

3.2. Autoencoders for Anomaly Detection

Autoencoders are a class of neural networks capable of reconstructing the given input at their output layer, with the constraint of deriving a low-dimensional representation at their middle layer (bottleneck) [31]. To be able to reconstruct well, an autoencoder must extract the most fundamental and abstract feature set within its bottleneck. As illustrated in figure 2, conventional autoencoders consist of two main blocks: the encoder – involving the extraction of the low-dimensional representation at the bottleneck – and the decoder, responsible for mapping latent space representation to the reconstructed input at the output layer.

Assuming x, z, $\hat{x}$, $f(.)$ and $g(.)$ as the original input, corresponding latent space variable, reconstructed input, encoder block, and decoder block data processes these two blocks can be formulated as $f(.) : x \rightarrow z$ and $g(.) : z \rightarrow \hat{x}$, respectively. The performance of an autoencoder is evaluated according to the similarity shared by the original input and the corresponding reconstructed version; therefore, functions comparative of the input and its reconstructed version are intrinsic choices of the loss function to train these networks; intuitively, more similarity between the original and reconstructed inputs indicates better performance of the autoencoder. The most usual

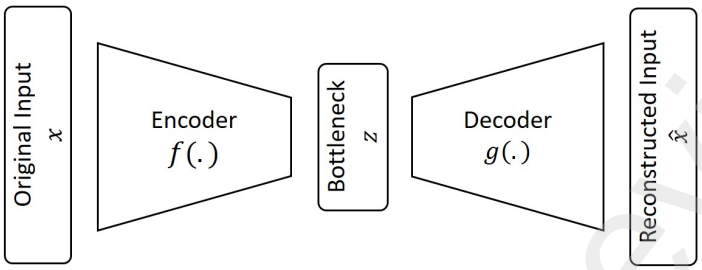


Figure 2: Visual Demonstration of the Schematic of an Autoencoder Network

choice of the autoencoder cost function is the mean-squared error (MSE), definable as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2, \tag{1}$$

where n is the length of the input. As mentioned in Section 2.2, the application of autoencoders for anomaly detection is well-established. Implementation of an autoencoder-powered anomaly detection solution starts with training an autoencoder utilizing normal-only information. Having access to the normal-only information during its training process, the autoencoder masters the reconstruction of the normal data by grasping its fundamental characteristics. For the case of anomalous data, however, it being unknown to the autoencoder during the training phase results in poor reconstruction performance. Since the relationships in the data change, the patterns the autoencoder relied on for bottleneck compression are no longer there. Hence, by comparing the reconstruction error (Rec_{err}) with a predefined threshold, it is possible to filter out the anomalous records; where Rec_{err} higher (lower) than the threshold are specified as anomalous (normal). If both normal and anomalous labeled data are available during the training stage, the threshold is easily determinable by direct comparison of the Rec_{err} over these two subsets. However, it is harder to determine the threshold in unsupervised anomaly detection scenarios. In these cases, statistical properties of Rec_{err}

have shown to allow discriminating normal from anomalous data (for example, different percentiles of error distribution over the training set are potential candidates to choose the threshold from [32]).

3.3. Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) is an explanation method capable of explaining neural networks, independent of the model input type (video, images, text, etc.) [33]. Unlike most of the local model-agnostic explainability methods where the reasoning behind a prediction is demonstrated in one step, LRP utilizes a backward pass through the model structure to assign a relevance score to each neuron in the network, from output all the way long back to the inputs [34]. It is worth mentioning that LRP redistributes the relevance score conservatively, meaning that the relevance score achieved by a neuron is equal to the summation of the relevance scores assigned to the neurons existing in the previous layer [33, 35]. This property of LRP is better understood by the examination of Equation 2, where i and j are two neurons located in consecutive layers (l and $l+1$), $R_i^{(l)}$ and ${R_j}^{(l+1)}$ are the relevance scores assigned to them and z_{ij} quantifies the contribution of neuron i to make neuron j relevant [35].

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} {R_j}^{(l+1)} \text{with } z_{ij} = {x_i}^{(l)} {w_{ij}}^{(l,l+1)} \tag{2}$$

The redistribution rule presented in Equation 2 is the basic rule available to use to compute the relevance scores; in addition to that, more advanced variants are developed to enhance the performance in special cases; examples are ϵ-variant to receive sparser explanations for the cases of weak or contradictory neurons and γ-variant which generates more stable explanations by asymmetric treatment of positive and negative contributions [33].

3.4. Similarity-based fault classification

The PIAFC method works based on the similarity shared between an arbitrary faulty signal and the expected fault patterns. To do so, we first generate a set of annotation vectors – one per fault type, and with the length of the original signal – for each rotating speed. These annotations highlight the neighboring frequency of the fault characteristic frequency. Let us take $A = [a_1, a_2, ..., a_i]$ to be the annotation matrix with the size of $n \times i$, where n, i, and a_i are the length of the original signal, the number of faults and

annotation vectors, respectively. Using the similarity evaluation property of the inner product between two vectors, we can assess the similarity of each annotation vector to an arbitrary signal. Based on that, the similarity score of the original signal and each annotation vector (Z vector of length i) is calculated according to Equation 3, where $X[n]$ is the original signal.

$$Z = X[n] \times A \tag{3}$$

Derived similarity scores lack the conservation property – their sum does not equal 1; to fix this, we use a softmax operation:

$$\sigma(Z)_k = \frac{e^{z_k}}{\sum_{j=1}^{i} e_j^z}$$

to map the similarity scores to a normalized vector, intuitively similar to probability logits.

Different types of profiles can be used to achieve such a highlighting pattern; moreover, different numbers of harmonics of the fault characteristic frequency may be used to build up the annotation vectors. In the case study used in this paper, four different profiles (square, triangular, parabolic, and hyperbolic) are compared. These profiles are demonstrated in Figure 3. It is worth mentioning that profiles within this figure are drawn using a 20 Hz neighborhood centered around 103.36 Hz. Similarly, our visual inspection of a set of observations from the case study dataset showed that harmonics higher than the third one are rare to find; thus, we utilize annotation vectors with the first three harmonics of the fault characteristic frequencies.

Last but not least, different trending patterns according to the value of the peaks at different harmonics in the annotation vectors are applicable. The intuition behind the variation of the value of the harmonics is to differentiate the amount of attention paid to different harmonics. In our study, we take into account three different ones:

- constant: same value for all the harmonics
- increasing: higher values for higher harmonics
- declining: lower values for higher harmonics

4. Experimental Evaluation

We start this section by introducing the datasets used in our case studies and the corresponding data preparation and preprocessing steps; next

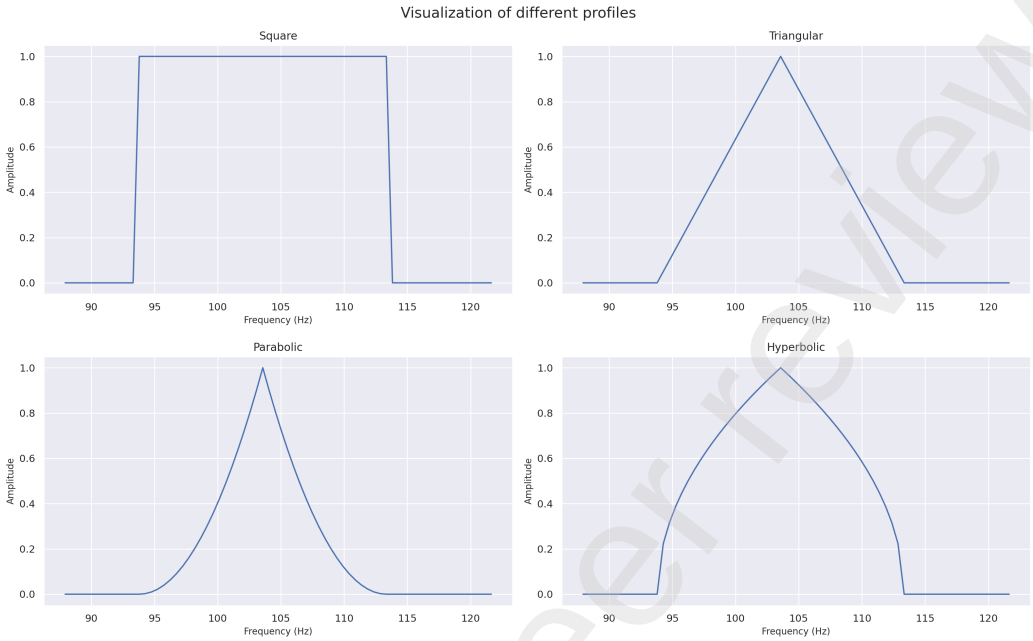


Figure 3: Visual demonstration of the annotation profiles

comes the introduction and discussion of both single-component and multi-component fault diagnosis case studies.

4.1. Introduction to Datasets

4.1.1. Case Western Reverse University (CWRU) bearing dataset

This dataset is among the most frequently used benchmark datasets of rotary machine fault diagnosis; mainly due to its richness of metadata and well-measured records. This dataset considers four different health states of rolling element bearings, including normal, inner-ring fault, outer-ring fault, and ball fault. Additionally, to consider the challenge of loading condition variation, acceleration of the bearing is recorded at four different rotational speeds, including 1730 RPM (28.83 Hz), 1750 RPM (29.16 Hz), 1772 RPM (29.53 Hz), and 1797 RPM (29.95 Hz). Moreover, measurements identical to three different levels of severity of faults, according to the depth of the single-pointed fault, are used in this study. We included all the loading conditions and all the fault severity levels mentioned earlier in this study. It is worth mentioning that the CWRU dataset includes both Fan-End and Drive-End

Table 1: Frequency fault components by rotational speed for CWRU dataset (Drive-End bearing)

Fault	Ratio	Fault Frequency Component by Rotational Speed (Hz)			
		28.83	29.16	29.53	29.95
Inner-Race	5.4152	156.14	157.94	159.93	162.19
Outer-Race	3.5848	103.36	104.56	105.87	107.36
Ball	4.7135	135.91	137.48	139.21	141.17

Table 2: Frequency fault components by rotational speed for CWRU dataset (Fan-End bearing)

Fault	Ratio	Fault Frequency Component by Rotational Speed (Hz)			
		28.83	29.16	29.53	29.95
Inner-Race	4.9469	142.64	144.28	146.10	148.16
Outer-Race	3.0530	88.03	89.05	90.17	91.44
Ball	3.9874	114.97	116.30	117.76	119.42

acceleration signals.

Traditionally, the occurrence of bearing faults is investigated by the examination of frequency domain signals and recognition of specific frequency components expected to appear. These frequency components are multiplications of the bearing rotational speed and ratios calculated according to the bearing geometry. In Table 1 and Table 2, ratios and fault frequencies expected to be examined in the CWRU dataset for Drive-End and Fan-End bearings are available, respectively[1].

4.1.2. Bearing dataset by Society for Machinery Failure Prevention Technology (MFPT)

MFPT bearing dataset[2] is one of the earliest examples of rotary machines fault diagnosis dataset; yet still one of the best available datasets mainly due to: 1) its noticeable quality – from a data collection perspective – and 2) richness of metadata and operating conditions. It consists of bearing vibra-

[1]Ratios presented in are taken from:
`https://engineering.case.edu/bearingdatacenter/bearing-information`

[2]Available at:
https://www.mfpt.org/fault-data-sets/

tion signals, including its normal operation, inner-race and outer-race faults. The MFPT dataset does not consider the rotating speed variation, however, vibration signals for different loads – from 25 to 270 or 300 lbs – are available for some scenarios. It is worth mentioning that the sampling frequency in this dataset is inconsistent for all the measurements made; however, this is not an issue as the Zoom FFT technique derives frequency-domain representations within the interested frequency range and with the desirable frequency resolution, independent of sampling frequency. Similar to the ratios provided in Table 1 and Table 2, one is able to calculate bearing fault characteristic frequencies. To do so, we use the formulas provided in [36] for **Ball Pass Frequency Inner Race** and **Ball Pass Frequency Outer Race** – BPFI and BPFO for short – as follows:

$$BPFI = \frac{b \times f}{2}[1 + \frac{d}{e} \times cos(\beta)] \tag{4}$$

$$BPFO = \frac{b \times f}{2}[1 - \frac{d}{e} \times cos(\beta)], \tag{5}$$

where f, b, d, e, and β are the rotational frequency, number of rolling elements, ball bearing diameter, bearing pith diameter, and the bearing contact angle, respectively. According to the values provided for these parameters, BPFI and BPFO for 25 Hz are obtained as 118.88 and 81.12 Hz.

4.2. Data preparation and Preprocessing

4.2.1. CWRU dataset

Data preprocessing starts with splitting the original time domain signals presented in the dataset into 2048 point-long segments. Following that, the Hilbert transform is used to extract the envelope signal out of the original time signal. Mostly due to severe modulation between bearing fault frequencies and the rotational speed of the bearing, bearing faults are not easily recognizable in the frequency spectrum of raw acceleration time signals. As illustrated in [36], using the envelope of the time signal rather than the original time signal to extract the frequency spectrum is a powerful technique to overcome the modulation issue. In Figure 4, original frequency spectra and envelope frequency spectra of identical time signals – for different faults of bearing at the rotational speed of 1730 RPM – are visualized. In these figures, the expected frequency components are indicated using the red dashed

line. As it is clear, frequency spectra derived from envelope signals are considerably more adaptable to what is expected. Next, we employed Zoom FFT to extract frequency spectra in these figures to better concentrate on the 0 to 1000 Hz frequency range, which bearing fault frequencies rely on. Lastly, Zoom FFT is employed using all the 2048 points available to achieve a 0.4885 Hz frequency resolution in the frequency band of 0 to 1000 Hz. It is noteworthy that we apply the Hann window and a Butterworth bandpass frequency filter – with the range of 15 Hz to 800 Hz and a degree of 25 – to encounter frequency leakage error and aliasing, before the Zoom FFT.

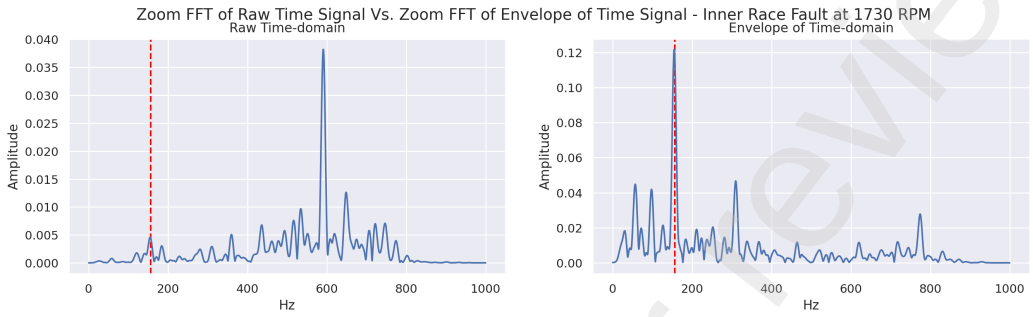
As we are dealing with an unsupervised anomaly detection task in this study, no faulty data is required during the training stage. Therefore, we used 75% of the normal data for training, and all the faulty data available, alongside the remaining 25% of the normal data, for evaluation purposes. Additionally, min/max scaling is used as the feature scaling, mainly since by the application of this method, every frequency component available in the frequency spectra is dealt with as a separate feature.

4.2.2. MFPT dataset

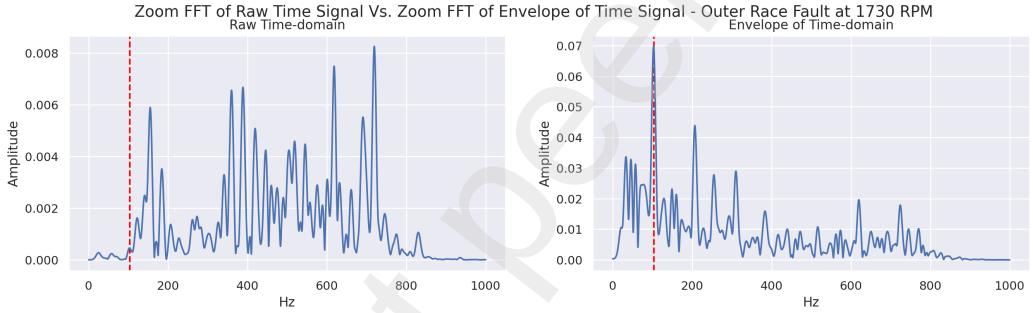
Most of the operations employed to preprocess the MFPT dataset are identical to the ones used for the CWRU dataset; however, as the sampling frequencies used within the MFPT dataset – 48828 and 97656 Hz – are noticeably different than the one utilized in CWRU – 12000 Hz – windowing lengths are different. To preserve the same amount of content within the windowed time domain signals, window length is chosen based on the sampling frequency. Therefore, signals with the sampling frequencies of 48828 and 97656 Hz are chosen to be 16671 and 8337, respectively. This way, the length of time that each measurement has virtually lasted is similar to the one with the CWRU dataset. The rest of the operations are identical to the ones used during the preprocessing of CWRU.

4.3. Single-component fault classification

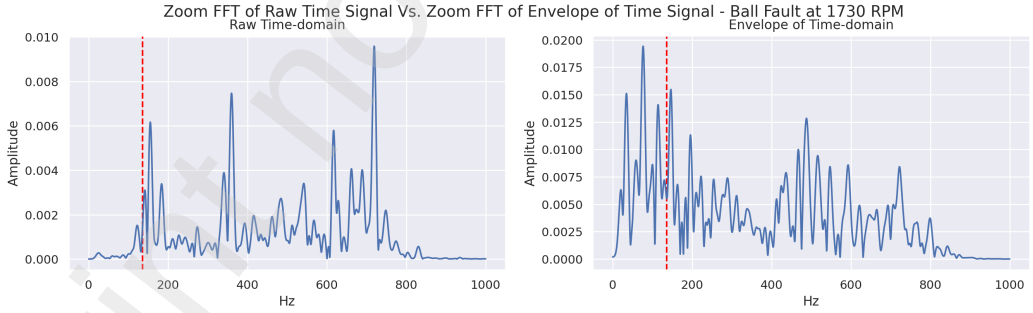
Anomaly Detection: The starting point to show the PIAFC pipeline used for single-component fault classification is to train the anomaly detection autoencoder. This autoencoder utilizes a symmetric multi-layered perceptron architecture, consisting of the following form as the number of neurons per layer: 2048-1024-512-256-512-1024-2048. Moreover, tanh, 0.0001, 3000, ADAM, and mean-squared error (MSE) are used as activation function, learning rate, number of epochs, optimizer, and loss function. The



(a) Original Frequency Spectrum Vs. Envelope Frequency Spectrum of Inner Race Fault



(b) Original Frequency Spectrum Vs. Envelope Frequency Spectrum of Outer Race Fault



(c) Original Frequency Spectrum Vs. Envelope Frequency Spectrum of Ball Fault

Figure 4: Visual Demonstration of the Proposed Method

Table 3: Mean and standard deviation of MSE by different health state for DE and FE bearings

State	Fan-End		Drive-End	
	mean	STD	mean	STD
Normal	0.0037	0.0012	0.0037	0.0001 <
Inner-race	13.8886	0.0213	2.1659	0.0062
Outer-race	23.2183	0.0383	7.3605	0.0190
Ball	75.8624	0.0171	1.0575	0.0042

chosen combination of learning rate and epochs ensures smooth and monotonic minimization of the proposed loss function, respectively. Additionally, to avoid overfitting, 25% of the training data is separated and used only for validation purposes. It is noteworthy that Pytorch is used for implementation purposes, and this autoencoder is trained on both Drive-End (DE) and Fan-End (FE) signals of the CWRU dataset, separately.

Table 3 summarizes the mean and standard deviation of the MSE (reconstruction error) by the health states over five trials. These values are derived from the held-out testing datasets of each bearing. The means of MSEs presented demonstrate that normal and anomalous data are well-separable by the threshold selection approach. As discussed in Section 3.2, for unlabeled training data, assumed to be normal, thresholds are specified using statistical features of the training reconstruction error. In this study, we consider maximum training reconstruction error ($m =$), its twice, its thrice, its quintuple, and the 95% of the training reconstruction error distribution as thresholds worth evaluating. Alongside that, we consider the threshold chosen as the decimal fractions around the mean of training $Rec_{err} = 0.0014$, namely 0.001, 0.005, 0.01, and 0.05. It is worth mentioning that large thresholds increase the number of false positives, while too small thresholds tend to increase the number of false negatives.

In Table 4, we summarize the binary classification accuracy of the autoencoder, with different threshold values; additionally, in Table 5, the same result is reported for several other unsupervised anomaly detection methods – including one-class SVM (OCSVM), isolation forest (IF) and their deep versions (Deep OCSVM) and (Deep IF) introduced in [37]. OCSVM and IF are among the best available choices to fulfill unsupervised anomaly detection problems; they are especially highly advantageous when it comes to hyperparameter tuning. Moreover, their deep versions involve the applica-

Table 4: Mean and standard deviation of binary classification accuracy (normal/anomalous) for autoencoder (m and p correspond to the maximum and the 95 percentile of the training reconstruction error, and values beneath them are identical to their averages for FE and DE bearings)

Bearing	Threshold Value								
	0.001	0.005	0.01	0.05	m (0.0027) (0.0027)	$2 \times m$	$3 \times m$	$5 \times m$	p (0.0021) (0.0019)
FE	0.9141 (< 0.0001)	0.9871 (0.0009)	**0.9996** (0.0002)	0.8902 (0.0012)	0.9344 (0.0031)	0.9916 (0.0009)	0.9993 (0.0002)	0.9964 (0.0003)	0.9160 (0.0003)
DE	0.9343 (<0.0001)	0.9936 (0.0006)	0.9978 (0.0001)	0.9513 (0.0011)	0.9441 (0.0019)	0.9960 (0.0011)	**0.9988** (0.0004)	0.9968 (0.0005)	0.9344 (0.0001)

tion of these methods on the feature space derived at the latent space of the autoencoder, instead of the original feature space. This is a counteraction to overcome the notorious **curse of dimensionality**: "*when the number of attributes or features increase, the amount of data needed to generalize accurately also grows, resulting in data sparsity in which data points are more scattered and isolated*" [38]. It is worth mentioning that OCSVM and IF are fit to the same data autoencoder is trained on, using the implementation and default hyperparameters provided by the scikit-learn[3] library.

According to the values presented in Tables 4 and 5, autoencoder with a proper threshold tends to outperform all the competing methods. Additionally, the autoencoder offers a higher level of performance consistency (according to the standard deviation of the accuracy) in comparison with other methods, except for the IF. Therefore, the autoencoder is certainly a reliable anomaly detection method. One noteworthy thread to pull is that although normal and faulty subsets of the test set are very far, according to the means provided for Rec_{err} in table 3, this does not result in perfect classification accuracy. We believe this originates in that, although the autoencoder generally performs poorly in reconstructing faulty subsets, a few cases are still reconstructed well, resulting in imperfect classification performance.

Fault Classification: Next comes the implementation of the physics-informed fault classification approach, introduced in Section 3.4. As a faulty signal must belong to one of the three cases of inner-race, outer-race, and ball faults, for a running rotational speed, an annotation matrix with the size of 2048×3 is defined. Once the matrix multiplication illustrated in Equation

[3]https://scikit-learn.org/stable/

Table 5: Mean and standard deviation of binary classification accuracy (normal/anomalous) for competing methods

Bearing	Method			
	IF	OCSVM	Deep IF	Deep OCSVM
FE	0.9555 (<0.0001)	**0.9954** (0.0016)	0.9911 (0.0034)	0.9695 (0.0009)
DE	0.9670 (<0.0001)	**0.9967** (0.0005)	0.9929 (0.0022)	0.9740 (0.0017)

3 and the softmax operator are applied, we obtain the probability-like logits capturing the class the faulty signal most likely belongs to.

As mentioned in Section 3.4, various combinations of the harmonic profiles and trending patterns are available. We summarized the classification accuracies of the DE bearing signals for various combinations in Table 6. Considering no need for training data, the proposed method performs acceptably as a passive (training-free) pseudo-labeling approach. In Figure 5, the confusion matrix of the best combination of table 6 (Declining-Parabolic) is presented; according to this figure, the proposed approach performs the weakest in the classification of the ball problem signals. This is justifiable since, according to the literature, the envelope fails to reveal fault characteristic frequency components of a ball fault class signal, cf. [39, 40]. It is worth mentioning that the poor performance of the proposed method in the classification of ball problem signals is not restricted to this special combination; in fact, this class possesses the highest number of misclassifications for all the combinations of trending patterns and harmonic profiles.

Next, we get into the application of XAI to explain the autoencoder reconstruction process. We chose LRP mainly due to two reasons: 1) LRP applies to any arbitrary deep learning method regardless of its architecture, unlike many XAI methods – e.g. Class Activation Maps – whose applications are limited to certain architectures, and 2) LRP is highly advantaged according to computational efficiency [34]. It is worth mentioning that we use the implementation provided by the Captum library to compute relevance scores [4].

LRP derives a relevance score for each neuron existing at the output layer,

[4] https://captum.ai/api/lrp.html

Table 6: Classification accuracy, for various combinations of the harmonics profiles and trending patterns (first, second and third are highlighted using **bold**, underlined and *italic*)

Trending pattern	Harmonic profile	Classification accuracy
Declining	Square	0.7296
	Triangular	0.7326
	Parabolic	**0.7352**
	Hyperbolic	0.7315
Constant	Square	0.7149
	Triangular	*0.7333*
	Parabolic	0.7344
	Hyperbolic	0.7273
Increasing	Square	0.6742
	Triangular	0.7019
	Parabolic	0.7146
	Hyperbolic	0.6852

which in this study is a relatively high dimensional space (consisting of 2048 neurons). Such a huge feature space makes inspecting the relevance scores assigned to the input for each output neuron unpractical. To overcome this issue, we use averaging to compute the mean contribution of every frequency component in the input, over all the neurons present in the output layer of the autoencoder. This way, we evaluate the mean contribution of every frequency component, in the reconstruction of the output. In figure 6, the intuition behind averaging the LRP values is visually demonstrated.

In Figure 7, we visualized one example per each DE bearing fault with its corresponding LRP values. In these figures, black dashed lines represent the first and second harmonics of the fault frequency components, provided in table 1. As it is obvious, in figure 7a and figure 7b significant increase in the LRP value content at the neighboring frequency range of the expected frequency components is sensible. For the ball fault scenario, however, neither the original acceleration signal nor the corresponding LRP values share high-quality semantic information regarding the expected frequency components; due to the inability of the HT to reveal the fault characteristic frequency components in the ball fault. To quantify the effectiveness of LRP values to correct misclassified records (by the similarity-based classifier), we examined all the misclassified signals and their corresponding LRP values together to categorize them into the following classes: **1) Confirmative:** confirming

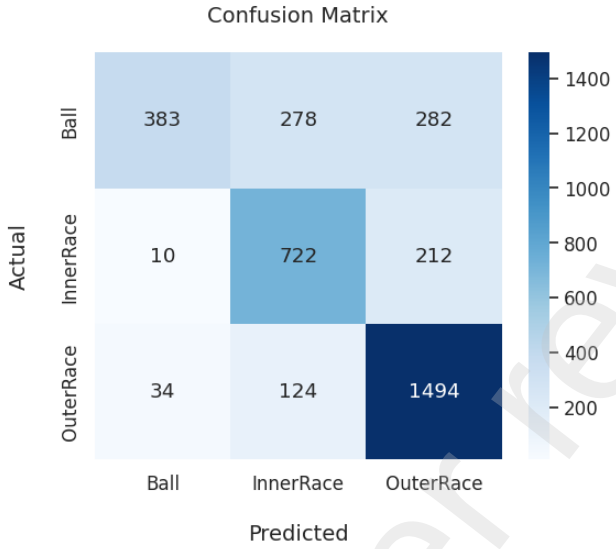


Figure 5: Confusion matrix of the physics-informed approach, for the case of declining and parabolic as trending pattern and harmonic profile

the wrong prediction, **2) Corrective:** Correcting the wrong prediction, **3) Neutral:** LRP values did not help, and **4) Too Noisy:** LRP values were too noisy to make a solid conclusion. Out of the 940 signals misclassified by the similarity-based classifier (declining-parabolic combination as the best combination of the table 6), we recognized 447, 237, 228 and 28 for classes 1 to 4; accordingly, in more than 25% of the misclassified records, a human expert can correct the prediction, using the LRP values alongside the original signal.

We suggest that readers try out the interactive dashboard available at the following link to get a comprehensive understanding of how our pipeline works; through this dashboard, users can analyze DE-bearing acceleration signals from different rotational speeds and health classes:
`https://demo-single-dashboard.onrender.com/`

In Figure 8, a screenshot is provided from this dashboard. As illustrated, the user interface consists of four main components; signal properties, where an arbitrary signal is chosen by setting a set of parameters; the graph of the original signal and its corresponding LRP values; annotation properties, where you choose the rotational speed and fault class you desire to check

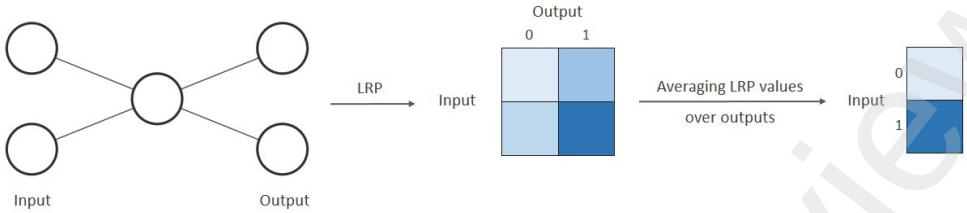


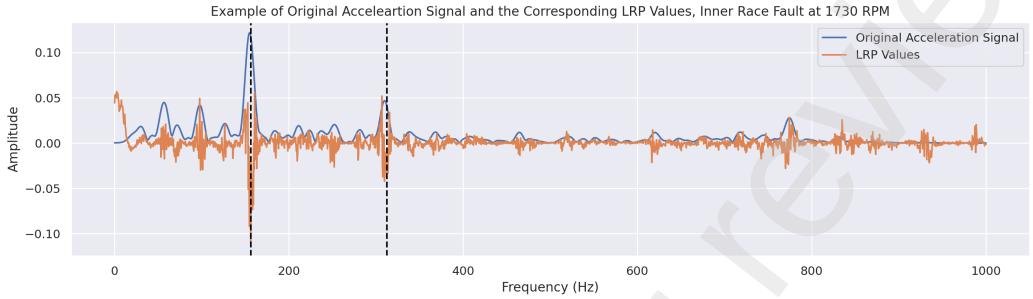
Figure 6: Visual demonstration of the LRP averaging technique

the vertical fault characteristic frequency component indicators; and additional information, where fault characteristic frequency and the normalized similarity score for the specified fault class are reported.

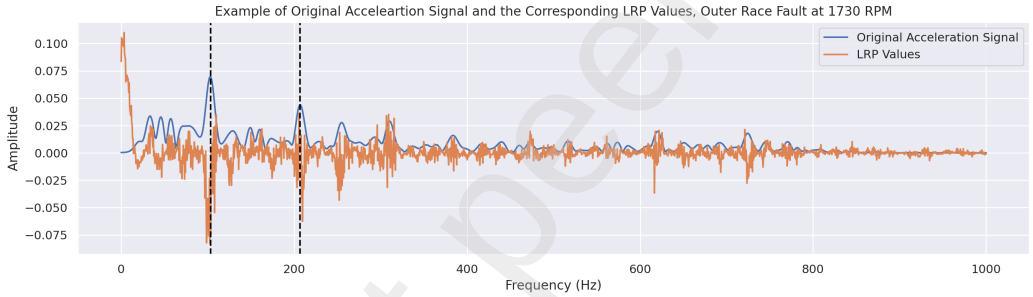
4.4. Multi-component fault classification

Anomaly Detection: In the previous section, we showed the efficacy of the proposed method on a single-component fault classification scenario; however, PIAFC is also applicable to fault detection and localization within multi-component setups. To demonstrate this, we use a case study involving the detection and localization of bearing faults from both FE and DE bearings of the rotating machinery test bench presented within the CWRU dataset. It is worth mentioning that we take both FE and DE signals available for this case, however, the pipeline is implementable utilizing merely one of them.

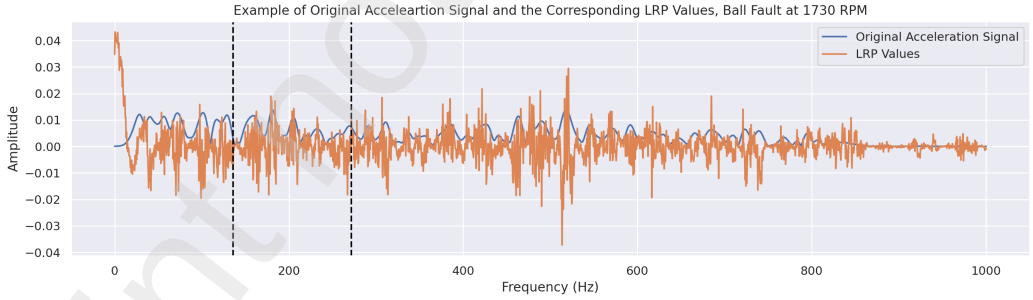
Again, the starting point is the implementation of the anomaly detection module. Unlike the single-component case where merely one signal is available, in this scenario, we need to adapt the anomaly detection module so it benefits from both FE and DE signals. To do so, one approach is to train an identical autoencoder on both FE and DE signals; an alternative is to train two separate autoencoders where each one is trained on signals coming from either FE or DE bearings. Table 7 summarizes the means and standard deviations of MSE over different combinations of DE and FE bearing subsets of the testing data for each of the approaches discussed above. Similar to the case of the single-component, the mean of MSE over the Normal-Normal state – which is the only absolutely normal state we are interested in as normal – is orders of magnitude lower than the faulty states and, therefore, easily separable by proper choice of the threshold; however, the dual model implementation achieve higher normal-anomalous separability, due to the higher



(a) Example of Envelope Spectrum Signal with Its LRP Values, Inner Race Fault at 1730 RPM



(b) Example of Envelope Spectrum Signal with Its LRP Values, Outer Race Fault at 1730 RPM



(c) Example of Envelope Spectrum Signal with Its LRP Values, Ball Fault at 1730 RPM

Figure 7: Examples of Envelope Spectrum Signals, with their Corresponding LRP Values for Different Bearing Faults, at 1730 RPM

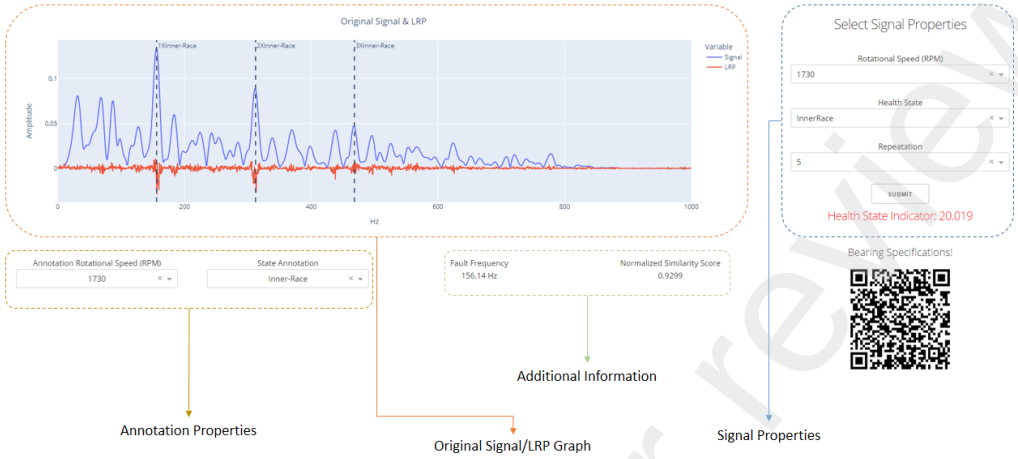


Figure 8: Screenshot of the single-component fault classification dashboard

difference of MSE means across normal-normal and anomalous subsets of the data. Additionally, the mean of MSE from the vibration signature taken from the normal bearings of faulty cases is significantly higher than the ones belonging to the Normal-Normal case, for both approaches. This proves the fact that this pipeline is implementable using single-point measurements too, rather than multi-point measurements. It is worth mentioning that, unlike the values presented in Table 3, these values are not identical to different iterations.

Fault Classification: Following the anomaly detection module, comes the adaptation of the physics-informed fault classification approach for the multi-component case. We do so, by the adjustment of the annotation matrix A as follows:

$$A = (a_{ir}^{FE}, a_{or}^{FE}, a_{b}^{FE}, a_{ir}^{DE}, a_{or}^{DE}, a_{b}^{DE}), \tag{6}$$

where an arbitrary a_u^v represents an annotation vector for the fault u and the bearing v. Each annotation vector is a column vector with the size of 2048×1, which makes the A a matrix with the size of 2048×6. Moreover, for each rotating speed, one A of this kind is definable. For implementation purposes,

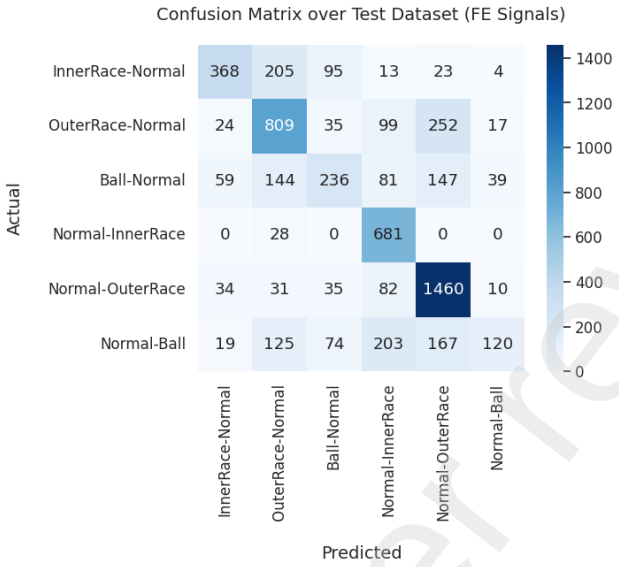
Table 7: Mean (and standard deviation, in parenthesis) of MSE over different health state setups. For every combination of bearing channels and states, the best performance is **bolded** (i.e., the highest mean of MSE, except for the *Normal/Normal* case, in which the lowest is best).

FE/DE State	FE		DE	
	Dual Model	Single Model	Dual Model	Single Model
Normal/Normal	0.0060 (0.0034)	**0.0020** (0.0008)	0.0048 (0.0017)	**0.0021** (0.0008)
Inner/Normal	**0.3909** (0.2443)	0.2958 (0.1974)	**2.3204** (1.9068)	1.7759 (1.5311)
Outer/Normal	**0.9998** (2.4137)	0.9042 (2.2588)	**7.3501** (12.2856)	6.6031 (11.3461)
Ball/Normal	**0.3155** (0.6893)	0.2380 (0.5457)	**1.0788** (2.1614)	0.8043 (1.6891)
Normal/Inner	**8.1624** (8.0557)	6.8923 (6.7767)	**0.5240** (0.5496)	0.4043 (0.4450)
Normal/Outer	**20.0168** (19.8403)	17.0205 (16.9483)	**0.6234** (0.5230)	0.4967 (0.4497)
Normal/Ball	**0.3827** (0.6100)	0.2896 (0.4977)	**0.0230** (0.0232)	0.0121 (0.0158)

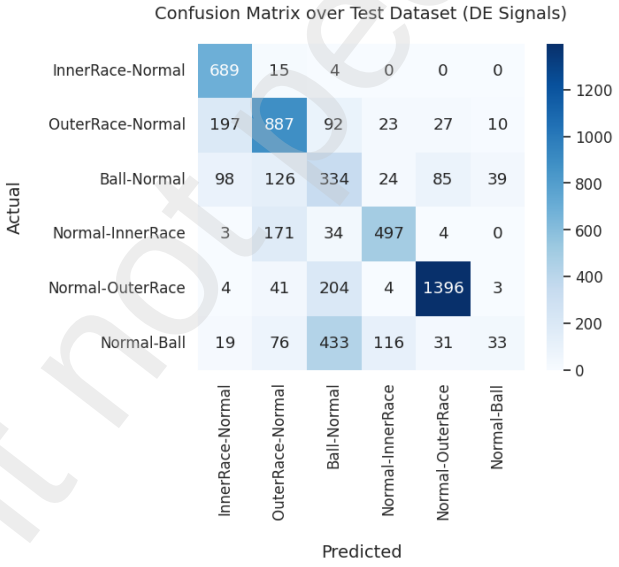
we decided to use the combinatory choices of declining and parabolic, as they outperformed other combinations in the previous section. The classification performance of the FE predictor and DE predictor are respectively 0.64242 and 0.6707; moreover, considering only the cases where both predictors share the same prediction, the classification accuracy is increased by 0.9165. It is worth mentioning that, two predictors are likely to agree in more than 60% (3510 out of 5719) cases of faulty observations. The significant improvement in classification accuracy shows that taking advantage of the multi-point measurement setup is definitely preferable. In Figure 9, confusion matrices describing the performance of these predictors are visualized; comparing the Figure 9c with Figures 9a and 9b, better performance of this approach when both bearing predictors agree is better shown. It is worth mentioning that classes in these illustrations are *FE bearing state + '-' + DE bearing state* (e.g. Normal-Ball means the cases where the FE bearing is normal and the ball problem is introduced to the DE bearing).

Following the physics-informed fault classification approach comes the XAI module for the multi-component setup. Similar to the single-component case, we also continue using LRP for multi-component cases. The only difference is that we now get a pair of explanations corresponding to the pair of measurements. Using these explanations, the frequency ranges driving each autoencoder to high MSE can be highlighted. In figure 10, examples of original envelope spectrum signals and their corresponding LRPs for the case of normal FE and faulty DE bearings are visualized. Similar to the plots of figure 7, the original signal, LRP values, and fault characteristics harmonics are colored blue, orange, and black, respectively. Similarly, various cases of faulty FE and normal DE bearings are also visualized in figure 11.

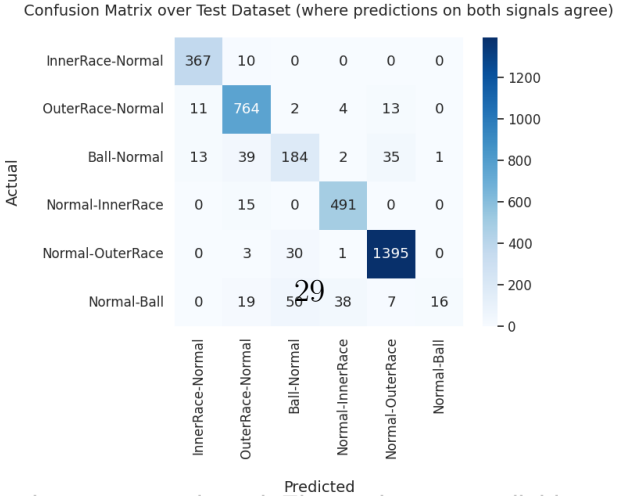
According to these two sets of figures, the following points are noteworthy: 1) fault signatures from each bearing are also traceable in the spectrum derived from the other bearing; therefore, the proposed pipeline is capable of multi-component fault detection employing merely measurements from a single source and 2) in some cases, fault characteristic peaks of issues in one component are more dominant and easier to recognize in the spectrum belonging to the other component (such as figure 10b, Figure 10c and Figure 11c). The latter phenomenon can be explained as follows: vibrations due to the fault at a bearing are capable of inducing secondary vibrations, resulting in a more disturbed spectrum and making the recognition of the primary source of vibration (bearing fault) harder; however, the secondary vibrations are not strong enough – in comparison with the primary source of vibration –



(a) Confusion matrix for the FE signal predictor



(b) Confusion matrix for the DE signal predictor



(c) Confusion matrix for the cases where both FE and DE signals predictors agree

Table 8: AE classification performance for different anomalous data leakage ratios (DE bearing)

Tr	Classification accuracy per anomalous data ratio					
	1	5	10	15	20	25
0.001	**0.9307**	**0.9307**	**0.9307**	**0.9307**	**0.9307**	**0.9307**
0.005	0.9933	0.9928	0.9933	0.9928	**0.9942**	0.9936
0.01	**0.9981**	0.9978	0.9978	**0.9981**	**0.9981**	0.9978
0.05	0.9508	**0.9519**	0.9508	0.9505	0.9497	0.9488
m	0.9385	0.9349	0.9344	0.9385	0.9385	**0.9419**
$2 \times m$	0.9944	0.9928	0.9930	0.9947	0.9950	**0.9964**
$3 \times m$	0.9989	0.9992	**0.9994**	0.9986	0.9986	0.9989
$5 \times m$	0.9972	**0.9975**	0.9972	0.9969	0.9972	0.9969
p	**0.9313**	0.9310	0.9307	0.9310	0.9310	0.9310

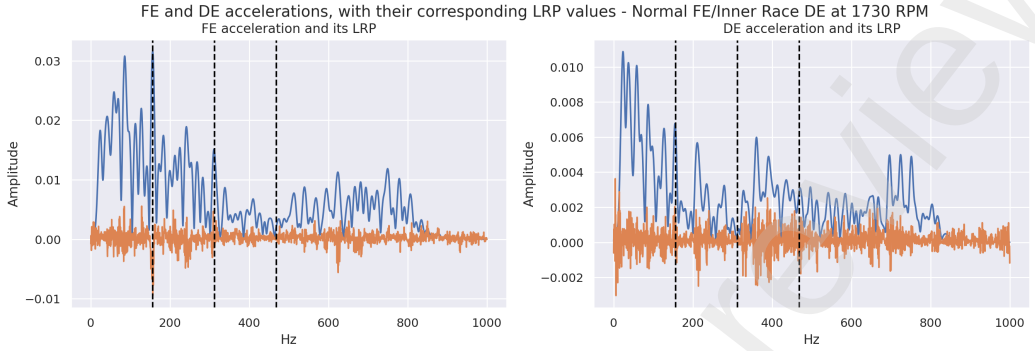
to preserve their high amplitude once they are emitted to the other bearing. Thus, fault characteristic components possess more dominant peaks, which are easier to recognize, at the frequency spectrum of the other bearing.

Similar to the single-component scenario, we have prepared a dashboard that enables readers to get a hands-on experience of how this approach works. The dashboard is available at:
`https://demo-dual-dashboard.onrender.com`

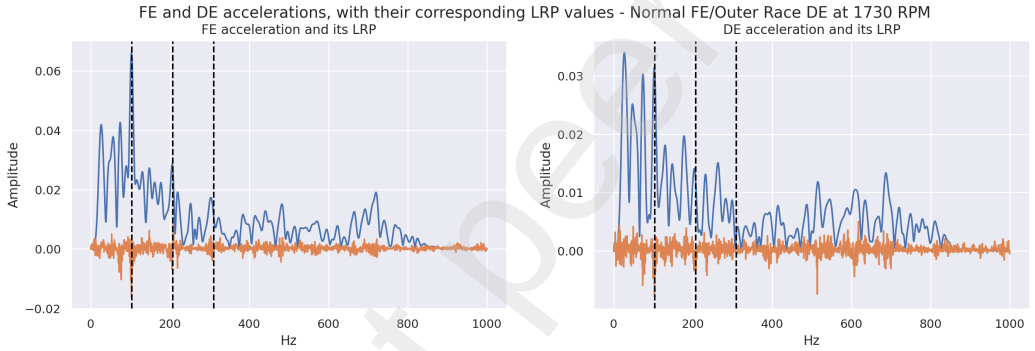
5. Ablation Studies

5.1. Robustness towards Poisonous Training

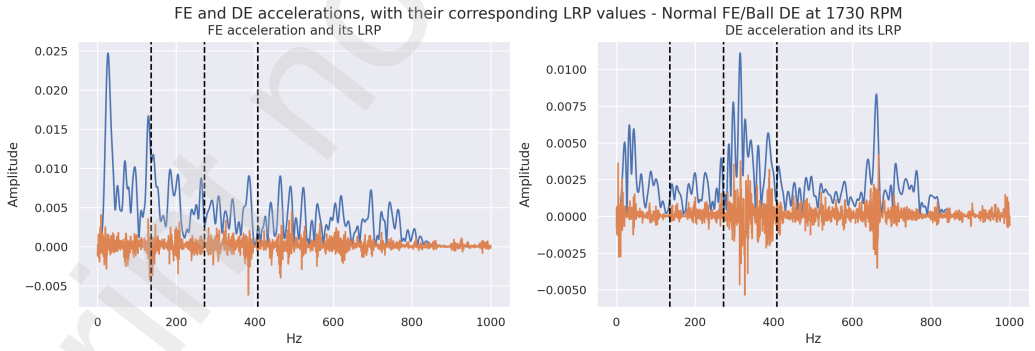
In experiments conducted in Section 3.4, we assumed that the training set only consists of normal data; however, this assumption is not always justified, especially for real-world industrial machinery data, which is likely to have anomalous data also present. To evaluate the robustness of AE towards the leakage of anomalous data to the training set, we conduct a set of experiments where the original fault-free training set is poisoned with different amounts of anomalous data. In table 8 and 9, classification accuracy corresponding to the poisonous training of the autoencoder on signals belonging to the DE bearing and FE bearing is summarized, respectively; comparing these tables with table 4 demonstrates robustness against poisoned training data.



(a) Normal FE and inner-race fault DE at 1730 RPM

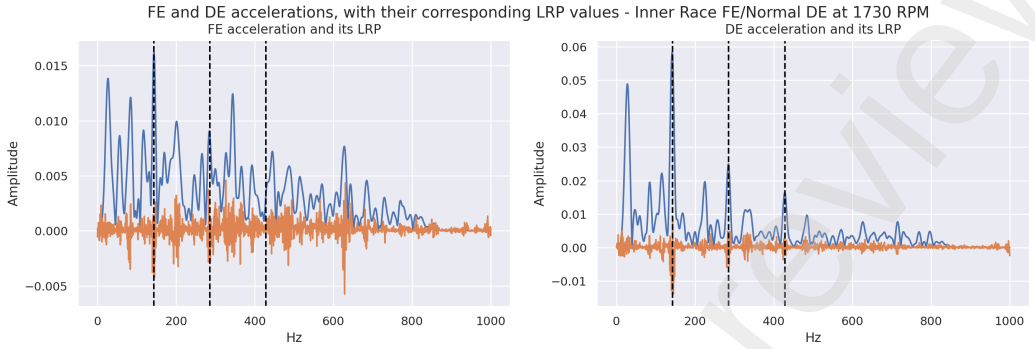


(b) Normal FE and outer-race fault DE at 1730 RPM

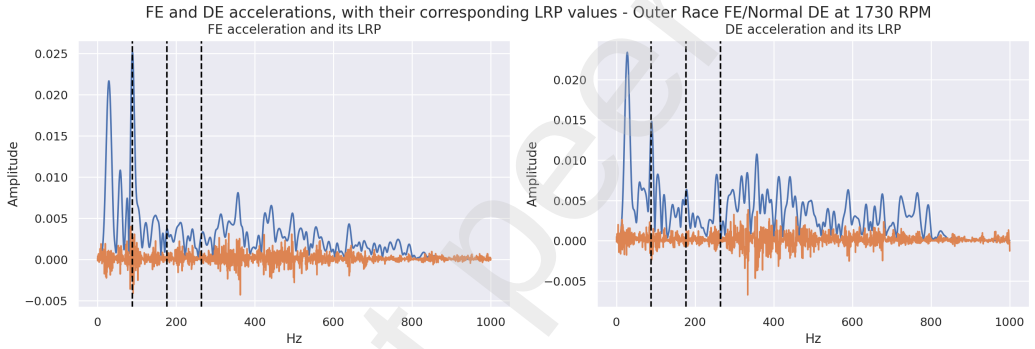


(c) Normal FE and ball fault DE at 1730 RPM

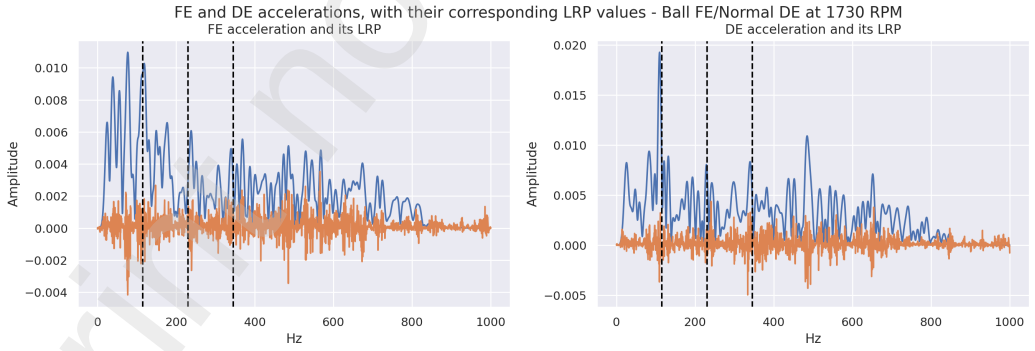
Figure 10: Examples of envelope spectrum signals with their LRP values for cases of normal FE and faulty DE, at 1730 RPM



(a) inner-race fault FE and Normal DE at 1730 RPM



(b) outer-race fault FE and Normal DE at 1730 RPM



(c) ball fault FE and Normal DE at 1730 RPM

Figure 11: Examples of envelope spectrum signals with their LRP values for cases of faulty FE and normal DE, at 1730 RPM

Table 9: AE classification performance for different anomalous data leakage ratios (FE bearing)

Tr	Classification accuracy per anomalous data ratio					
	1	5	10	15	20	25
0.001	**0.9080**	**0.9080**	**0.9080**	**0.9080**	**0.9080**	**0.9080**
0.005	0.9860	0.9856	0.9856	0.9856	**0.9867**	0.9856
0.01	**0.9996**	**0.9996**	0.9993	0.9993	0.9993	**0.9996**
0.05	0.8902	0.8914	0.8880	0.8910	**0.8917**	0.8910
m	0.9261	0.9276	**0.9298**	0.9268	0.9250	0.9231
$2 \times m$	0.9893	**0.9900**	**0.9900**	0.9897	0.9897	0.9871
$3 \times m$	0.9989	**0.9993**	**0.9993**	0.9989	0.9989	0.9985
$5 \times m$	0.9970	0.9963	0.9959	0.9967	0.9963	**0.9974**
p	0.9095	0.9098	**0.9106**	0.9098	0.9098	**0.9106**

5.2. Evaluation of PIAFC for transfer learning

Transfer learning (TL) aims to improve the performance at a target domain by taking advantage of previously learned knowledge from a source domain [41]. Recently, great attention has been paid to the TL applications for intelligent fault diagnosis; mainly as using TL approaches to unlock higher performance levels, target domain training data is limited [3]. This is done by the assumption that knowledge from one diagnosis task use case is not only reusable but also beneficial to solve a fairly similar diagnosis task [42, 3]. In this section, we evaluate how helpful the introduced method within a TL-based fault diagnosis approach is. To do so, we investigate a TL case study, where both source and target components are the same equipment type (bearing) but different specific assets. Thus, our main interest is to estimate the generalizability of what had been learned from the diagnosis of one bearing to an unseen one.

To do so, we designed a setup where the DE vibration signals from the CWRU dataset are regarded as the source domain and the MFPT dataset – introduced in section 4.1.2 – as the target domain. The following points are worth mentioning: 1) the MFPT dataset is preprocessed as discussed in section 4.2.2 and 2) a min/max scaler from the source domain is used to normalize the signals of the target domain. In Table 10, the reconstruction performance of a model trained on the source domain (DE bearing vibrations of the CWRU dataset) over each health class of the target domain

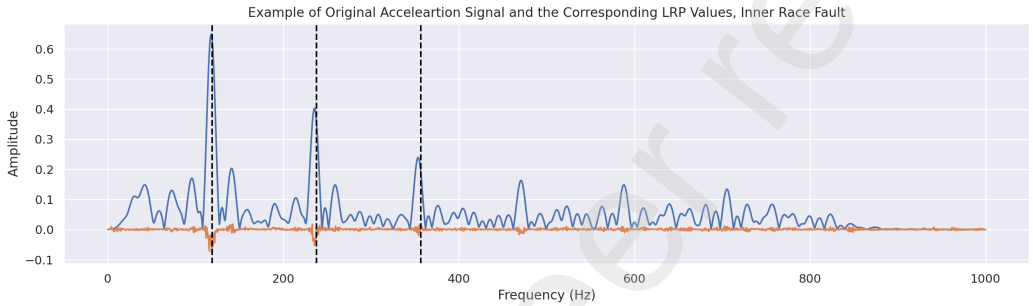
Table 10: Mean and standard deviation of MSE losses, over different health states of the MFPT dataset

Bearing health class	Mean of MSE	STD of MSE
Normal	3.7326	0.5380
Outer race fault	420.9286	128.0277
Inner race fault	38.3882	64.9731

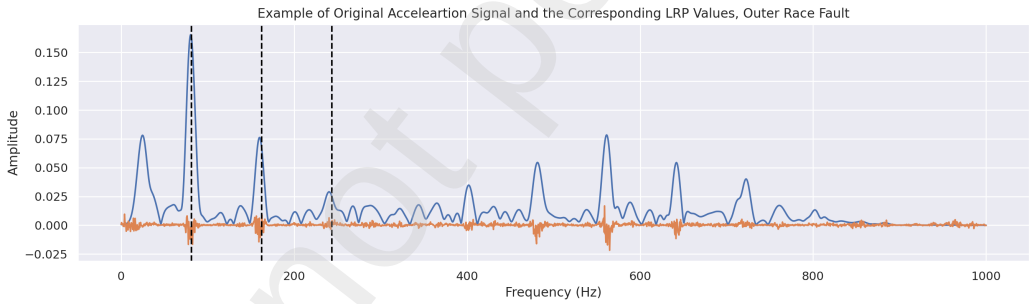
(MFPT dataset) is summarized. According to this table, while the normal and anomalous subsets of the target domain dataset are easily separable due to significant differences in the order of magnitude of mean MSE loss over them; however, the gap between the mean MSE loss of the source and target domains is huge and needs to be taken care of if the MSE threshold from the source domain is preferred to be used.

Following the evaluation of the anomaly detection autoencoder according to the reconstruction loss metric comes the examination of the LRP values to investigate if the sensitivity to the fault characteristic frequencies is preserved in the target domain. In figure 12, examples of the observations from the target domain and their corresponding LRP values are visualized. Similar to the previous figures, the coloring scheme is orange, blue, and dashed black for the original signal, corresponding LRP values, and harmonics of the fault characteristic frequencies. According to these figures, the autoencoder is sufficiently sensitive to the frequency ranges, including the harmonics of the fault characteristic frequencies; therefore, the proposed approach offers great potential to diagnose unseen pieces of machinery in a TL scenario, once the performance gap over normal data from source and target domains is taken care of.

Within the literature of TL, the performance drop between the source and target domains is due to **domain shift**: a significant difference in data distribution, between the source and target domains. By analyzing examples of signals and their corresponding LRP values from the normal class of the MFPT dataset, we can specify the frequency ranges driving the autoencoder to higher MSE losses. In Figure 13, we demonstrate such a case; the dominant peak at 25 Hz – the rotational speed – accompanied by intensified LRP values at the left side neighborhood of 25 Hz. Based on the fault signature presented in this figure – dominant peak at the rotational speed – the domain shift seems to be due to a fault with **1× peaks** as fault signature, e.g., shaft unbalancing, eccentricity, or a fault similar to those; nevertheless, once this



(a) inner-race fault at 25 Hz



(b) outer-race fault at 25 Hz

Figure 12: Examples of envelope spectrum signals with their LRP values for cases of faulty bearing, at 25 Hz

domain shift is taken care of, reconstruction performance over the normal data from the target domain is supposed to be adapted to the one expected of normal data from the source domain.

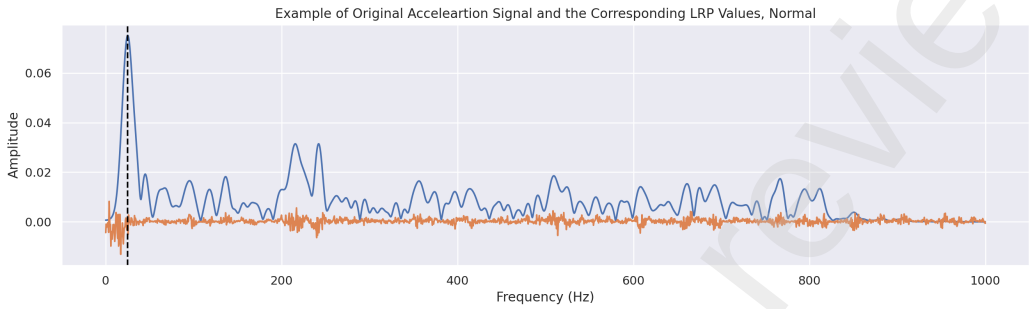


Figure 13: Example of envelope spectrum signal with its LRP values for normal bearing, at 25 Hz

5.3. Benchmarking XAI methods

The crucial role played by XAI methods within the proposed method brings up the importance of a performance comparison between different choices of XAI methods. This section is aimed to provide a comparison between a set of XAI methods. Two metrics are of interest: (1) meaningfulness of the explanations, according to the physically expected patterns, and (2) computational cost of the generation of those explanations, inference time of the XAI method, in other words, for different batch sizes. Moreover, the following points were of great importance in choosing between the available options:

- methods for the deep learning paradigms are preferred, therefore, SHAP and similar frequently used methods out of this scope are not included.
- defining a baseline reference for models manipulating vibration signals – anomaly detection tasks specifically – is an uncharted area to explore; therefore, models with such requirements, such as Integrated Gradients [43], are avoided.
- methods with specific architectural properties, e.g., Grad-CAM [44], are also avoided, due to loss of generalization.

It is worth mentioning that all the employed XAI methods have experimented using the implementations provided by the Captum[5] library.

5.3.1. Brief introduction of the methods of interest

The following are the XAI methods in which, their performance is compared with the previously introduced LRP:

Input X Gradient or IxG for short is a domestic explainability method for deep learning models, where the contribution of each neuron is assessed by differentiating its activation with its reference activation [45]. Calculation of the explanations in this method involves deriving the signed partial derivative of the output concerning each input and multiplying it with the input itself [46].

DeepLift compares the activation of each arbitrary neuron with its reference value and quantifies the difference made by the variation of the input, to the output and its reference value [47]. Similar to the previous method, here we also have the option to multiply the contribution scores with the input, in an element-wise manner. Moreover, the inclusion of references or baselines is not mandatory; as stated in the Camptum implementation documentation, *"In the cases when baselines are not provided, we internally use zero scalar corresponding to each input tensor"*[6].

Guided Backpropagation is an extension to **deconvnet** [48], which is used to *"project the feature activations back to the input pixel space"* [49]. More specifically, **Guided Backpropagation** takes into an additional constraint to zero out the gradients in comparison with the **deconvnet** – where only neurons with negative backward gradients are filtered out; it also zero masks neurons whose forward values are negative [50].

5.3.2. Comparison of inference time

We start by comparing the computation time of different methods over different sizes of batches to estimate the computational cost of the proposed pipeline if each of these methods was embodied in the XAI module. To keep the setup fair, we execute all the experiments in this section using a consistent computational resource, a Google Colab pro session, powered by T4 GPU and 25 GB of RAM. Moreover, to get a more stable version of the computational cost, these experiments are repeated 5 times, and mean and

[5]https://captum.ai/

[6]https://captum.ai/api/deep_lift.html

Table 11: Average (standard deviation) of execution time, for batch size (for each batch size, the lowest average of execution time is **highlighted**)

Method	Average (standard deviation) of execution time, for batch size						
	1	5	10	50	100	500	1000
LRP	10.3925 (0.8983)	10.5663 (0.3289)	10.5516 (0.4050)	12.7246 (0.8095)	14.8436 (0.4521)	30.9907 (2.3450)	55.4165 (1.9839)
IxG	4.6442 (1.0657)	**4.7955** (1.0935)	**4.5633** (0.1780)	**6.3921** (0.1606)	**8.3222** (0.3267)	**25.6784** (1.3839)	**48.4996** (2.0086)
Guided Backprop	**4.4566** (1.0949)	5.1911 (1.2921)	4.8545 (0.5120)	6.8281 (0.4346)	8.7994 (0.2707)	26.7299 (1.6939)	50.9672 (1.9767)
DeepLift	29.2925 (1.8822)	28.2254 (1.2647)	28.7669 (0.6865)	31.1274 (0.1382)	34.2302 (0.8817)	61.5091 (2.5538)	96.7022 (3.2510)

standard deviations are summarized in table11. To better compare these methods, values presented in this table are also visualized in figure 14.

According to the values presented in the table above IxG and Guided Backprop are the main two methods competing for the lowest computation time over all the batch sizes experimented. Moreover, LRP performs moderately among all the methods included in this study, according to the average computation time, however, it offers better scalability performance for batch sizes from 50 to higher. Moreover, the DeepLift method performs the worst, according to the average computation cost. Additionally, IxG shows the lowest standard deviation for the batch sizes from 10 to 500 – and it is slightly outperformed by LRP and Guided Backprop at the batch size of 1000; in other words, this method performs most consistently in this range of batch sizes. This property is of great importance, as we are considering the application of this step as a constant subroutine of a data processing approach.

5.3.3. Semantic comparison of explanations

Unlike comparing the computation time, we carried out in the previous section using a quantitative fashion, comparing the explanations provided by the methods above for the semantically meaningful insights is done in a qualitative approach. To do so, in figure 15, examples from different faults at 1730 RPM and corresponding explanations from the XAI methods are illustrated.

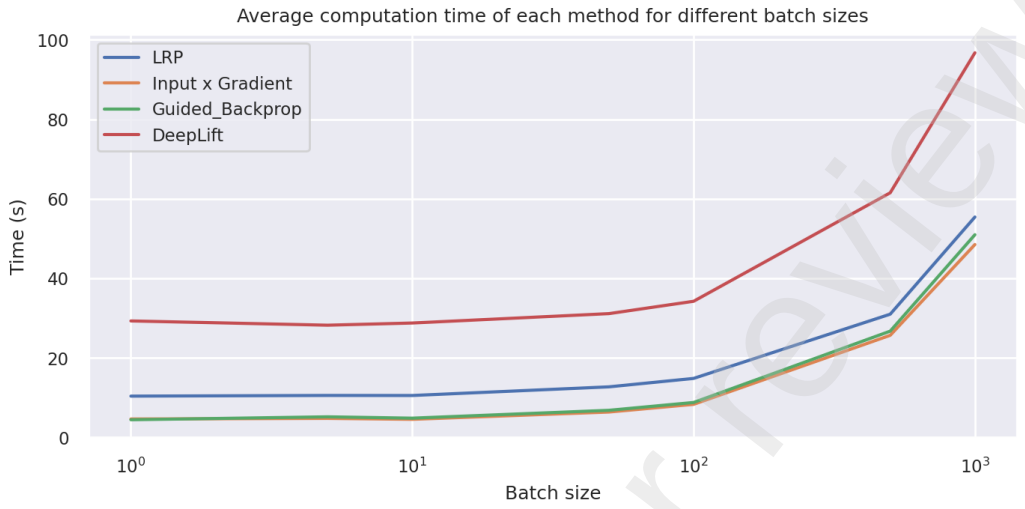


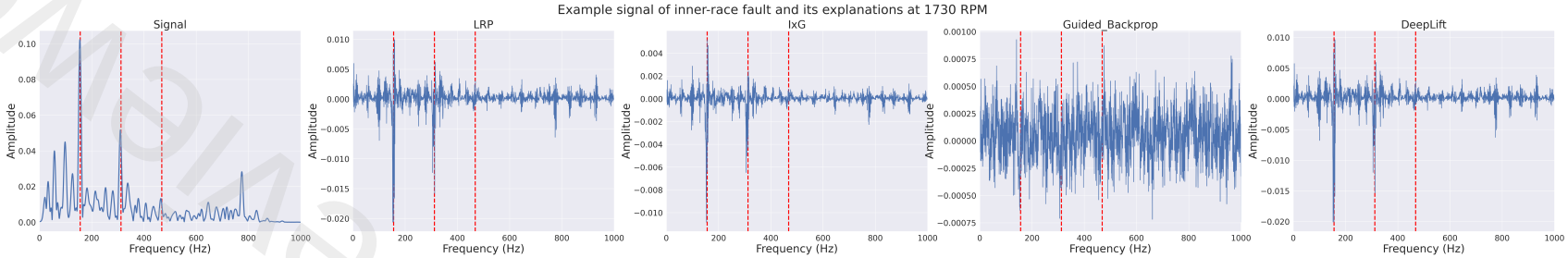
Figure 14: The average computation time of each XAI method for different batch sizes

According to these figures, the explanations derived by the Guided Backprop method do not convey any sort of semantically meaningful insight, comparatively. Moreover, explanations provided by LRP and DeepLift show almost identical explanations; however, their computational cost according to the values presented in table 11 differ significantly. Last but not least, IxG also performs neatly in showing sensitivity to the harmonics of the fault characteristic component, but the absolute magnitudes of the negative peaks are noticeably smaller in amplitude, even orders of magnitude in some cases.

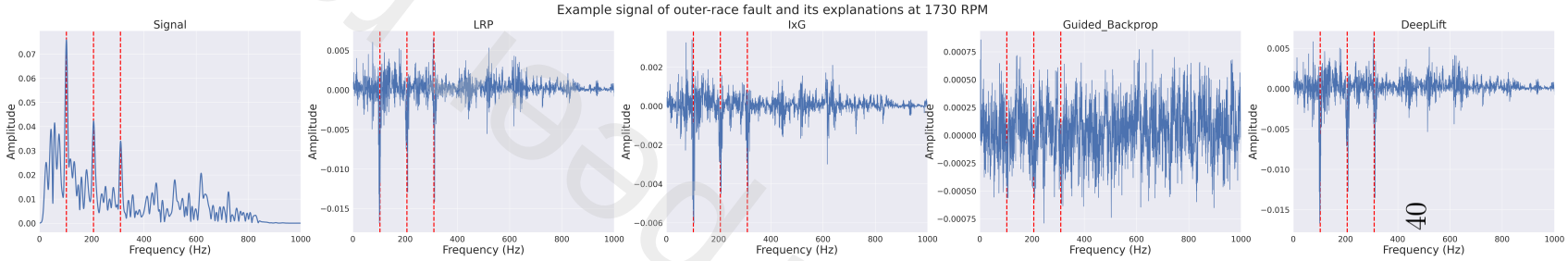
6. Conclusions

In this study, we introduced PIAFC; a hybrid approach based on XAI and component-level physical understanding that, in addition to the preliminary anomaly detection task, is capable of recognizing the most probable failure mode. PIAFC offers several advantages: the most unprecedented one is that it is implementable using fault-free datasets. Moreover, its transparent nature makes it absolutely understandable to human experts. Last but not least, PIAFC offers impressive multi-component performance.

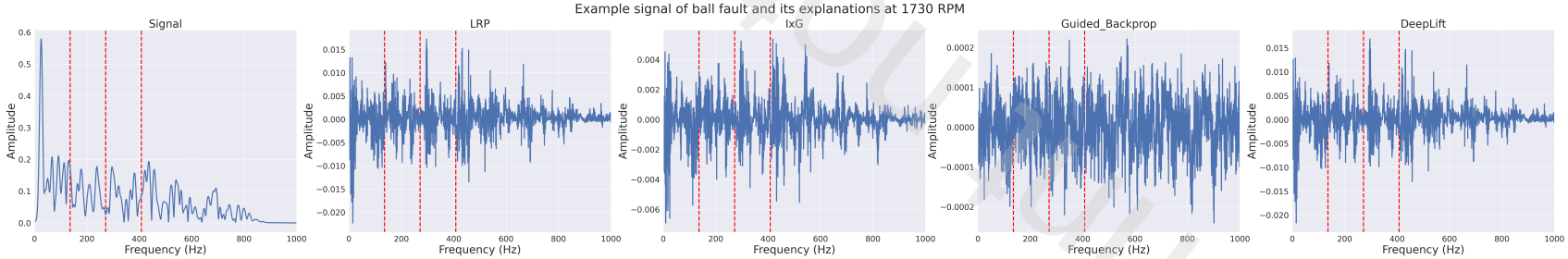
Experiments we present in Sections 4.3 and 4.4, demonstrate the applicability of PIAFC for single-component and multi-component fault classification scenarios, respectively. Additionally, throughout Section 5, we evaluate



(a) Example of inner-race fault at 1730 RPM with explanations from different XAI methods



(b) Example of outer-race fault at 1730 RPM with explanations from different XAI methods



(c) Example of ball fault at 1730 RPM with explanations from different XAI methods

Figure 15: Examples of different faults at 1730 RPM with explanations from different XAI methods

different properties of PIAFC; we start by investigating the robustness of PIAFC towards poisonous training in 5.1, showing that the binary classification accuracy of anomaly detection task is fairly robust to the anomalous data leakage to the training set; in fact, cases are present in Table 9 and Table 8, where the inclusion of the anomalous data in training set improves the classification accuracy. Next, in Section 5.2, we show that PIAFC has the potential to be used for transfer learning scenarios, as it is sensitive to faults from unseen components under unseen working conditions. Finally, in Section 5.3, due to the integral role of XAI in our method, we compare the performance of different XAI methods according to the inference time and semantic content. Nevertheless, PIAFC suffers from two limitations: first and foremost, the fusion of physical understanding and XAI requires human supervision, and therefore it is not possible to deploy PIAFC without human interaction. Second, knowledge of the operating working conditions and component properties is essential to analyze the fault patterns.

The current study can be extended in various directions. We believe that the highest priority is to apply PIAFC for new pieces of machinery, out of the rotating machines scope. The main requisite to take advantage of PIAFC is the physical understanding of machinery failure modes; relevant examples are power transformers (Duval triangle [51] to interpret the dissolved-gas-in-oil analysis, DGA), compressors (performance curves analysis for fault diagnosis purposes) and solar panels (maximum power point, MMP, analysis [52]). Another idea is to use indirect measurement methods (similar to what is done in [53, 54]) to monitor working conditions; this way, no additional sensors and acquisition hardware are required. Last but not least, similarity-based fault classification can replace human experts to label unlabeled vibration signals; using it alongside learning from noisy labels [55] (to encounter noisy labels produced by similarity-based fault classification method) strategies, near perfect classification performance will be achievable, without human labeling burden.

References

[1] H. Bendjama, S. Bouhouche, M. S. Boucherit, Application of wavelet transform for fault diagnosis in rotating machinery, International Journal of Machine Learning and Computing 2 (1) (2012) 82–87.

[2] J. Chen, C. Lin, D. Peng, H. Ge, Fault diagnosis of rotating machinery: a review and bibliometric analysis, IEEE Access 8 (2020) 224985–225003.

[3] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A. K. Nandi, Applications of machine learning to machine fault diagnosis: A review and roadmap, Mechanical Systems and Signal Processing 138 (2020) 106587.

[4] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, R. X. Gao, Waveletkernelnet: An interpretable deep neural network for industrial intelligent diagnosis, IEEE Transactions on Systems, Man, and Cybernetics: Systems 52 (4) (2021) 2302–2312.

[5] H. Lu, V. P. Nemani, V. Barzegar, C. Allen, C. Hu, S. Laflamme, S. Sarkar, A. T. Zimmerman, A physics-informed feature weighting method for bearing fault diagnostics, Mechanical Systems and Signal Processing 191 (2023) 110171.

[6] S. Shen, H. Lu, M. Sadoughi, C. Hu, V. Nemani, A. Thelen, K. Webster, M. Darr, J. Sidon, S. Kenny, A physics-informed deep learning approach for bearing fault detection, Engineering Applications of Artificial Intelligence 103 (2021) 104295.

[7] A. Rai, S. H. Upadhyay, A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings, Tribology International 96 (2016) 289–306.

[8] S. Zhang, F. Ye, B. Wang, T. G. Habetler, Few-shot bearing fault diagnosis based on model-agnostic meta-learning, IEEE Transactions on Industry Applications 57 (5) (2021) 4754–4764.

[9] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR, 2017, pp. 1126–1135.

[10] C. Li, S. Li, A. Zhang, Q. He, Z. Liao, J. Hu, Meta-learning for few-shot bearing fault diagnosis under complex working conditions, Neurocomputing 439 (2021) 197–211.

[11] S. Wang, D. Wang, D. Kong, J. Wang, W. Li, S. Zhou, Few-shot rolling bearing fault diagnosis with metric-based meta learning, Sensors 20 (22) (2020) 6437.

[12] A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, J. Hu, Limited data rolling bearing fault diagnosis with few-shot learning, Ieee Access 7 (2019) 110895–110904.

[13] Y. Zhang, S. Li, A. Zhang, C. Li, L. Qiu, A novel bearing fault diagnosis method based on few-shot transfer learning across different datasets, Entropy 24 (9) (2022) 1295.

[14] A. Berenji, Z. Taghiyarrenani, A. Rohani Bastami, Fault identification with limited labeled data, Journal of Vibration and Control 30 (7-8) (2024) 1502–1510.

[15] R. B. Amir, S. T. Gul, A. Q. Khan, A comparative analysis of classical and one class svm classifiers for machine fault detection using vibration signals, in: 2016 International Conference on Emerging Technologies (ICET), 2016, pp. 1–6. doi:10.1109/ICET.2016.7813262.

[16] S. Liu, Z. Ji, Y. Wang, Improving anomaly detection fusion method of rotating machinery based on ann and isolation forest, in: 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2020, pp. 581–584. doi:10.1109/CVIDL51233.2020.00-23.

[17] S. Ahmad, K. Styp-Rekowski, S. Nedelkoski, O. Kao, Autoencoder-based condition monitoring and anomaly detection method for rotating machines, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 4093–4102.

[18] M. Kaji, J. Parvizian, H. W. van de Venn, Constructing a reliable health indicator for bearings using convolutional autoencoder and continuous wavelet transform, Applied Sciences 10 (24) (2020) 8948.

[19] A. Matsui, S. Asahi, S. Tamura, S. Hayamizu, R. Isashi, A. Furukawa, T. Naitou, Anomaly detection in mechanical vibration using combination of signal processing and autoencoder, Journal of Signal Processing 24 (4) (2020) 203–206.

[20] F. Arellano-Espitia, M. Delgado-Prieto, V. Martínez-Viol, Á. Fernández-Sobrino, R. A. Osornio-Rios, Anomaly detection in electromechanical systems by means of deep-autoencoder, in: 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE, 2021, pp. 01–06.

[21] F. König, C. Sous, A. Ouald Chaib, G. Jacobs, Machine learning based anomaly detection and classification of acoustic emission events for wear monitoring in sliding bearing systems, Tribology International 155 (2021) 106811. doi:https://doi.org/10.1016/j.triboint.2020.106811.
URL https://www.sciencedirect.com/science/article/pii/S0301679X20306368

[22] G. Lee, M. Jung, M. Song, J. Choo, Unsupervised anomaly detection of the gas turbine operation via convolutional auto-encoder, in: 2020 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2020, pp. 1–6.

[23] M. Sadoughi, C. Hu, Physics-based convolutional neural network for fault diagnosis of rolling element bearings, IEEE Sensors Journal 19 (11) (2019) 4181–4192. doi:10.1109/JSEN.2019.2898634.

[24] Y. A. Yucesan, F. A. Viana, A physics-informed neural network for wind turbine main bearing fatigue, International Journal of Prognostics and Health Management 11 (1) (2020).

[25] Y. A. Yucesan, F. A. Viana, Hybrid physics-informed neural networks for main bearing fatigue prognosis with visual grease inspection, Computers in Industry 125 (2021) 103386.

[26] H.-Y. Chen, C.-H. Lee, Vibration signals analysis by explainable artificial intelligence (xai) approach: Application on bearing faults diagnosis, IEEE Access 8 (2020) 134246–134256.

[27] M. J. Hasan, M. Sohaib, J.-M. Kim, An explainable ai-based fault diagnosis model for bearings, Sensors 21 (12) (2021) 4070.

[28] M. S. Kim, J. P. Yun, P. Park, Deep learning-based explainable fault diagnosis model with an individually grouped 1-d convolution for three-axis vibration signals, IEEE Transactions on Industrial Informatics 18 (12) (2022) 8807–8817. doi:10.1109/TII.2022.3147828.

[29] L. C. Brito, G. A. Susto, J. N. Brito, M. A. Duarte, An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery, Mechanical Systems and Signal Processing 163 (2022) 108105.

[30] M. Kim, J. Yun, P. Park, An explainable neural network for fault diagnosis with a frequency activation map, IEEE Access PP (2021) 1–1. doi:10.1109/ACCESS.2021.3095565.

[31] M. Xia, T. Li, L. Liu, L. Xu, C. W. de Silva, Intelligent fault diagnosis approach with unsupervised feature learning by stacked denoising autoencoder, IET Science, Measurement & Technology 11 (6) (2017) 687–695. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-smt.2016.0423, doi:https://doi.org/10.1049/iet-smt.2016.0423.
URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-smt.2016.0423

[32] M. A. Salahuddin, M. F. Bari, H. A. Alameddine, V. Pourahmadi, R. Boutaba, Time-based anomaly detection using autoencoder, in: 2020 16th International Conference on Network and Service Management (CNSM), IEEE, 2020, pp. 1–9.

[33] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 193–209.

[34] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable ai methods-a brief overview, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 13–38.

[35] A. Binder, S. Bach, G. Montavon, K.-R. Müller, W. Samek, Layer-wise relevance propagation for deep neural network architectures, in: Information science and applications (ICISA) 2016, Springer, 2016, pp. 913–922.

[36] E. Bechhoefer, A quick introduction to bearing envelope analysis, Green Power Monit. Syst (2016).

[37] K. Sohn, C.-L. Li, J. Yoon, M. Jin, T. Pfister, Learning and evaluating representations for deep one-class classification, arXiv preprint arXiv:2011.02578 (2020).

[38] S. Thudumu, P. Branch, J. Jin, J. Singh, A comprehensive survey of anomaly detection techniques for high dimensional big data, Journal of Big Data 7 (2020) 1–30.

[39] A. Berenji, S. Nowaczyk, Z. Taghiyarrenani, Data-centric perspective on explainability versus performance trade-off, in: B. Crémilleux, S. Hess, S. Nijssen (Eds.), Advances in Intelligent Data Analysis XXI, Springer Nature Switzerland, Cham, 2023, pp. 42–54.

[40] M. Sohaib, C.-H. Kim, J.-M. Kim, A hybrid feature model and deep-learning-based bearing fault diagnosis, Sensors 17 (12) (2017). doi:10.3390/s17122876.
URL https://www.mdpi.com/1424-8220/17/12/2876

[41] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.

[42] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359. doi:10.1109/TKDE.2009.191.

[43] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks (2017). arXiv:1703.01365.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[45] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences (2017). arXiv:1605.01713.

[46] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Gradient-Based Attribution Methods, Springer International Publishing, Cham, 2019, pp. 169–191. doi:10.1007/978-3-030-28954-6_9.
URL https://doi.org/10.1007/978-3-030-28954-6_9

[47] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences (2019). arXiv:1704.02685.

[48] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Advances in neural information processing systems 31 (2018).

[49] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[50] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net (2015). arXiv:1412.6806.

[51] M. Duval, Fault gases formed in oil-filled breathing ehv power transformers-interpretation of gas-analysis data, in: IEEE Transactions on Power Apparatus and Systems, no. 6, IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC 345 E 47TH ST, NEW YORK, NY . . . , 1974, pp. 1745–1746.

[52] B. Nehme, N. K. Msirdi, A. Namaane, T. Akiki, Analysis and characterization of faults in pv panels, Energy Procedia 111 (2017) 1020–1029.

[53] J. Urbanek, T. Barszcz, J. Antoni, A two-step procedure for estimation of instantaneous rotational speed with large fluctuations, Mechanical Systems and Signal Processing 38 (1) (2013) 96–102.

[54] A. Beering, K.-L. Krieger, Estimation of rotational speeds based on gearbox vibrations via artificial neural networks, IEEE Sensors Journal 22 (23) (2022) 22963–22971.

[55] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE transactions on neural networks and learning systems 34 (11) (2022) 8135–8153.