Data-Centric Perspective on Explainability versus Performance Trade-off

Amirhossein Berenji[0000−0003−3720−3015], Sławomir Nowaczyk[0000−0002−7796−5201], and Zahra Taghiyarrenani[0000−0002−1759−8593]

Center for Applied Intelligence Systems Research, Halmstad University, Sweden
{firstname.lastname}@hh.se

Abstract. The performance versus interpretability trade-off has been well-established in the literature for many years in the context of machine learning models. This paper demonstrates its twin, namely the data-centric performance versus interpretability trade-off. In a case study of bearing fault diagnosis, we found that substituting the original acceleration signal with a demodulated version offers a higher level of interpretability, but it comes at the cost of significantly lower classification performance. We demonstrate these results on two different datasets and across four different machine learning algorithms. Our results suggest that "there is no free lunch," i.e., the contradictory relationship between interpretability and performance should be considered earlier in the analysis process than it is typically done in the literature today; in other words, already in the preprocessing and feature extraction step.

Keywords: Explainable AI · SHAP · Intelligent Fault Diagnosis · Bearings · Hilbert Transform · Envelope Spectrum

1 Introduction

Rotary machines are one of the most crucial pieces of equipment in industrial production [9]; they consist of a huge number of components, including bearings. Even non-severe bearing faults disrupt the normal operation of rotating machines. Bearing fault is also among the frequent failure modes of rotary machines; 40% to 50% of all failures in rotating machinery are estimated to be due to bearing faults [20]. Therefore, bearing condition monitoring is of great importance.

The promising performance of pattern recognition techniques in machine condition monitoring use cases resulted in the creation of Intelligent Fault Diagnosis (IFD) – the application of artificial intelligence methods for machine fault diagnosis [13]. Although IFD-based solutions often achieve super-human performance in scientific settings, their application in the industrial sector is relatively limited due to a lack of transparency. Therefore, the employment of eXplainable Artificial Intelligence (XAI) methods to provide insight into their reasoning is of high priority.

Over the last decades, the interpretability versus performance trade-off from the *model perspective* – i.e., the fact that higher performance is often associated with higher complexity, and thus usually achieved by sacrificing the interpretability – has been well established [5]. While improved interpretability is not necessarily followed by reduced model performance, maintaining the latter while improving the former typically requires conscious effort, and often advanced techniques [21]. In this study, we pose a complementary question “does the application of preprocessing methods to make IFD pipelines more interpretable necessarily degrade their performance?”

The contribution of this work is to bring attention to an inherent decrease in classification performance caused by replacing the original data with an interpretable representation. A bearing fault diagnosis case study with and without counter-modulation transformation is an example of such a situation. To compare the original data versus an interpretable version of it, we evaluate two different preprocessing branches. One includes the Hilbert Transform as a demodulation technique, while the other excludes it. As pointed out by [2], bearing faults are easier to recognize – for a human expert – in the frequency spectrum of a demodulated signal. The classification accuracy achieved by the two branches, however, shows the opposite effect. Comparing the performance of the two representations clearly demonstrates that, for an artificial neural network, such human-interpretable features are subpar compared to raw data.

The rest of the paper is organized as follows: we first investigate relevant earlier work in Section 2. Afterward, in Section 3, a brief scientific background of the employed methods is provided. Next, in Section 4, the experimental setup is explained in detail, while the corresponding results are discussed in 5. Finally, in Section 6, we provide a discussion of the findings and conclude the paper.

2 Related Works

Explainability is on its way to becoming a must in IFD implementations. For example, in [3], authors introduced an unsupervised classification approach based on the attribution of explainability from an anomaly detection model. The effectiveness of this method is evaluated not only by the application of different models but also by an examination of different datasets. The authors took advantage of Shapely Additive Explanations (SHAP) to derive the feature importance scores. Similarly, in [19], authors evaluated the effectiveness of different XAI methods, including Gradient Class Activation Map (Grad-CAM), Layer-wise Relevance Propagation (LRP), and Local Interpretable Model-agnostic Explanations (LIME), to explain a shaft imbalance detection model. Another approach is to incorporate physics-inspired features, cf [6]. Authors applied a Frequency-RPM transformation to transform time domain signals to time-frequency representation; these representations are usually regarded as images, and therefore Convolutional Neural Networks (CNNs) are widely applied to manipulate these representations. Lastly, in [4], Grad-CAM is applied to derive explanations from a CNN model used to diagnose bearing faults. Short-Time Fourier Transform (STFT) is used to extract the time-frequency representation of time-domain bearing acceleration signals. As the authors ignored the modulation phenomena in bearings, their derived explanations are not in good accordance with patterns expected physically; however, the authors then showed that patterns corresponding to different health states are repeatable and comparative.

Hilbert transform is frequently used to demodulate time domain signals. For example, in [11], authors used Hilbert transform for envelope extraction purposes, alongside cyclo-stationary analysis (to cope with non-stationary signals) to reveal fault frequency components in an air conditioning production assembly line. Moreover, Hilbert Transform is frequently used as the demodulation technique in bearing vibration analysis pipelines. As an example, in [16], authors used it alongside wavelet packet decomposition to extract the fault characteristics from the bearing acceleration signal. Similarly, authors of [22] showcased the effectiveness of the application of envelope analysis to reveal fault frequency components expected to observe in the Case Western Reverse University bearing dataset.

3 Background

3.1 Zoom FFT

Zoom FFT is a technique to improve frequency resolution within a specific frequency range [12]. Application of Zoom FFT not only reduces the length of the original signal to achieve the desired frequency resolution but also decreases the computational cost significantly [12]. Implementation of Zoom FFT consists of two main stages; the first one is the application of a group of operations to preprocess the original signal, while the second stage is the application of conventional FFT.

As illustrated in Figure 1, the preprocessing stage starts with a multiplication of the original signal ($x[k]$ with a length of N) by the complex signal of $[\cos(2\pi f_c t)+i\sin(2\pi f_c t)]$, where f_c is the lower limit of the desired frequency range ($[f_c, f_c+B_p]$). It continues with low-pass filtering of the multiplication signal (x_{mu}), using the bandwidth of B_p. Afterward, the filtered signal is undersampled by M (known as decimation), resulting in a signal with the length of N/M. Next, zero padding is employed to fill in for the $N-(N/M)$ instances removed during the decimation process. Finally, the FFT is employed to derive a frequency domain signal, within the desired frequency range, out of the zero-padded signal.

3.2 Hilbert Transform to Extract Envelopes

Hilbert Transform (HT) of a signal is defined [7] as:

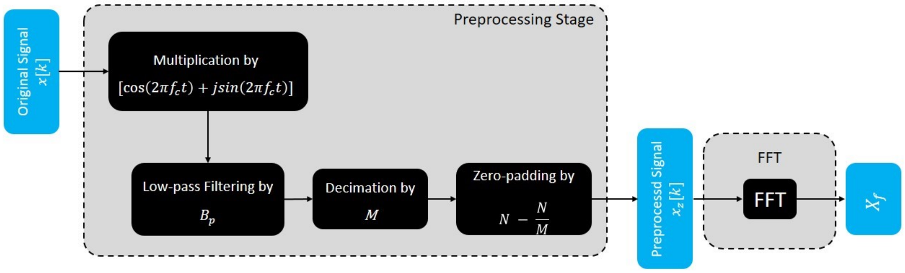

Original Signal $x[k]$
Preprocessing Stage
Multiplication by $[\cos(2\pi f_c t) + j\sin(2\pi f_c t)]$
Low-pass Filtering by B_p
Decimation by M
Zero-padding by $N - \frac{N}{M}$
Preprocessd Signal $x_z[k]$
FFT
FFT
X_f

Fig. 1: Visual illustration of Zoom FFT

$$H[x(t)] = \tilde{x}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(t)}{t-\tau} d\tau \tag{1}$$

Therefore, we can define an analytic signal as a complex function in which the real part is the original signal, and its imaginary part is the HT [7]:

$$X(t) = x(t) + i\tilde{x}(t), \tag{2}$$

where $X(t)$ is the analytic signal, $x(t)$ is the original signal, and the $\tilde{x}(t)$ is the HT of the original signal. Similar to any other time variant complex function, the instantaneous amplitude of the analytic signal can be computed as:

$$A(t) = |X(t)| = \sqrt{x^2(t) + \tilde{x}^2(t)} \tag{3}$$

The instantaneous amplitude of the analytic signal varies slower than the original signal [7]. Therefore, the instantaneous amplitude function – also known as envelope – is a version of the original signal excluding high-frequency oscillations. Accordingly, the envelope extraction based on HT is considered a demodulation approach widely used in rotating machinery vibration analysis [8].

4 Experiments

4.1 Introduction to Datasets

Most of our experiments are done on the Case Western Reverse University (CWRU) bearing dataset; it includes four different bearing health states: normal, inner-race fault, outer-race fault, and ball problems. We focus our study on Drive-End (DE) bearings, as DE bearings are subjected to more mechanical stresses in real-world scenarios. Signals with 48000 and 12000 Hz sampling frequencies are available; however, we found 12000 Hz sufficient. In this dataset, four levels of rotational speeds (1730 RPM, 1750 RPM, 1772 RPM, and 1797 RPM) are included, and we used them all to consider the challenge of variation in mechanical loading. The rotational speed is vitally important for bearing fault detection, as the occurrence of faults in the bearings is likely to exhibit dominant peaks at particular frequency components (fault characteristic components). These components are the multiplication of geometrically defined ratios by the rotational speed of the bearing. In Table 1, ratios of different faults[1] alongside the fault frequency component by the rotational speed are summarized.

Unfortunately, due to the modulation phenomena, the expected bearing fault components are not usually observable in frequency spectra; therefore, a demodulation step is essential to reveal the true fault frequency components.

To generalize our findings beyond a single dataset, we confirm our observations also using the Paderborn University (PU) bearing dataset [14]. We again focus on bearing fault classification, including normal and synthetically generated faults of the inner race and outer race. Moreover, we also considered mechanical loading variation by including both 900 and 1500 RPM shaft rotational speeds.

[1] Ratios from https://engineering.case.edu/bearingdatacenter/bearing-information

4.2 Data Preparation and Preprocessing

To study the effect of the application of HT on classification accuracy, we consider two preprocessing branches. Both preprocessing branches start with the initial step of splitting the original time domain signals to 2048 and 12800 points-long signals for CWRU and PU datasets, respectively. Following that, **on the first branch**, *raw*, we use a generally accepted pipeline for rotating machinery vibration analysis [23, 15, 18]. It starts with the application of a Hann window to avoid leakage error, and a Butterworth bandpass frequency filter (with a degree of 25 and cut-off frequencies of 2.5 Hz and 5500 Hz for CWRU, and 2.5 and 31000 for PU) is employed to both remove the DC components and prevent aliasing. Afterward, we applied Fast Fourier Transform (FFT) algorithm to derive the frequency spectrum. The resulting frequency domain signals are 1024 points-long signals, covering 0 to 6000 Hz and 0 to 32000 Hz for CWRU and PU datasets, respectively.

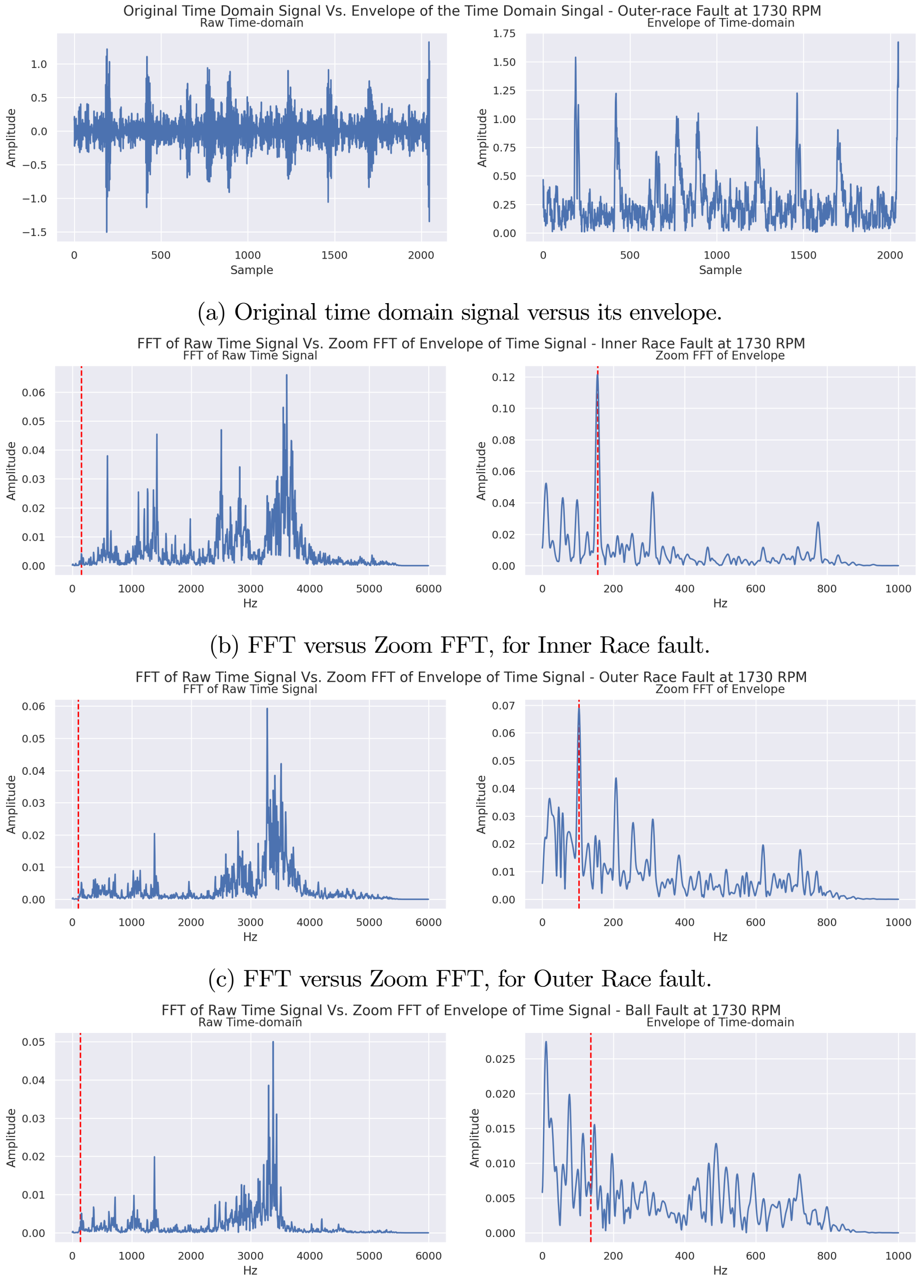
On the second branch, *envelope*, we take advantage of HT to extract the envelope from the raw time domain signal. Therefore, to derive a well-suited frequency resolution within the desired frequency range (0 to 1000 Hz), Zoom FFT is employed. The choice of the frequency range is made to cover not only the frequency components corresponding to the faults but also their initial harmonics. Moreover, since 1024 points are used to apply the Zoom FFT technique, the resulting frequency domain signals are also 1024 points long. Similar to the raw branch, we also used the Butterworth bandpass frequency filter prior to the application of Zoom FFT; however, the second cut-off frequency is 800 Hz.

In Figure 2a, an example of the original time domain signal and its envelope is visualized. A comparison of the two indicates that the application of HT is indeed capable of reducing the disturbance level in the time domain signal. Moreover, plots in Figures 2b, 2c and 2d show that the envelope preprocessing branch is more powerful in revealing characteristic frequency components for bearing faults. It is worth noting that the red dashed lines in these plots highlight the expected fault frequency component, according to the values presented in Table 1. It is also to be noted that all the plots visualized in Figure 2 come from the CWRU dataset; however, the insights from the PU dataset are analogous.

For the experiments, data is split so that 40% is the hold-out testing dataset, and 25% of the remaining data is used for validation purposes. Additionally, we employ min/max scaling to transform values of all the frequency components within the frequency spectra to the range from zero to one.

Table 1: Frequency Fault Components by Rotational Speed for CWRU Dataset

Fault	Ratio	Fault Frequency Component by Rotational Speed			
		1730 RPM	1750 RPM	1772 RPM	1797 RPM
Inner-Race	5.4152	156.14 HZ	157.94 Hz	159.93 Hz	162.19 Hz
Outer-Race	3.5848	103.36 HZ	104.56 Hz	105.87 Hz	107.36 Hz
Ball	4.7135	135.91 HZ	137.48 Hz	139.21 Hz	141.17 Hz



(a) Original time domain signal versus its envelope.

(b) FFT versus Zoom FFT, for Inner Race fault.

(c) FFT versus Zoom FFT, for Outer Race fault.

(d) FFT versus Zoom FFT, for Ball fault.

Fig. 2: Visual demonstration of the signals from each preprocessing branch

4.3 Training Classifiers

Our experiments start with the application of Multi-Layered Perceptrons (MLP) to classify signals from the CWRU dataset. Networks to classify signals from both preprocessing branches utilize the structure of 1024-512-256-128-64-4 as neurons per layer. For the training of the network on the data from the first preprocessing branch (the one with the application of FFT on raw time domain signals), a combination of 10^{-4} and 50 as the learning rate and the number of epochs, respectively, provides monotonic and smooth minimization of the categorical cross-entropy loss. Notably, the proposed architecture achieves repeatable 100% classification accuracy on the held-out test dataset. On the other hand, we experienced strong overfitting when training the same network on the second preprocessing branch (using envelope extraction and Zoom FFT). Our experiments showed that the highest classification accuracy is achieved using a learning rate of 10^{-5} and 150 epochs at the verge of overfitting. Nevertheless, perfect performance is not attainable anymore.

Additionally, to strengthen the claim of the ubiquity of the tradeoff and demonstrate that the difference in the performance of the two preprocessing branches is independent of the classification method and not specific to deep neural networks, we also trained a group of classic machine learning models – including Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) – utilizing data belonging to both preprocessing branches, on data from CWRU dataset. It is worth mentioning that all the hyper-parameters of these models were set to the default values of scikit-learn[2] library.

Finally, to generalize our findings beyond a single dataset, we decided to evaluate the classification performance of both preprocessing branches on the PU dataset. For the conventional preprocessing, we employed an MLP with the structure of 6400-2000-250-3 with the 10^{-5} and 200 as the learning rate and epochs, respectively. Similarly, for signals from the interpretable branch, the structure is 1024-256-64-3, and a learning rate of 10^{-5} with 250 epochs were utilized.

5 Results

Table 2 summarizes the classification performance of different methods for both datasets and preprocessing branches. We repeat each experiment 5 times to minimize the randomness effect of training. To be able to examine the misclassified observations one by one, we keep the train and test sets fixed across all the trials. Based on the results in this Table, the performance decrease caused by the substitution of *Raw FFT* data with the *Zoom FFT* is consistently seen for essentially all cases. The one exception is the DT's results on the PU dataset; however, since the performance of this method is overall very poor (barely any learning is done, and the result is essentially random), we do not consider this to be contradicting our claim.

To better understand the performance versus interpretability tradeoff showcased here, we analyze the observations consistently misclassified across all 5 trials. As presented in the rightmost columns of Table 2, for the MLP row on the CWRU

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

Table 2: Classification performance of different methods on both datasets and two preprocessing branches, over 5 trials (the "C" column denotes the number of consistently misclassified observations).

Dataset	Method	Preprocessing	Classification Accuracy			# Misclassified		
			Min	Avg	Max	Min	C	Max
CWRU	DT	Raw FFT	0.9491	0.9525	0.9565	76	0	89
		ZoomFFT on Env	0.9376	0.9482	0.9605	69	5	109
	RF	Raw FFT	0.9977	0.9982	0.9989	2	2	4
		ZoomFFT on Env	0.9851	0.9859	0.9874	22	9	26
	SVM	Raw FFT	0.9994	0.9999	1.0000	0	0	1
		ZoomFFT on Env	0.9468	0.9469	0.9473	92	84	93
	MLP	Raw FFT	1.0000	1.0000	1.0000	0	0	0
		ZoomFFT on Env	0.9760	0.9769	0.9788	37	35	42
PU	DT	Raw FFT	0.7395	0.7629	0.7816	378	4	451
		ZoomFFT on Env	0.7556	0.7839	0.8018	343	26	423
	RF	Raw FFT	0.8914	0.8951	0.8983	176	54	188
		ZoomFFT on Env	0.8862	0.8889	0.8925	186	91	200
	SVM	Raw FFT	0.9041	0.9074	0.9110	154	135	166
		ZoomFFT on Env	0.8723	0.8776	0.8833	202	142	221
	MLP	Raw FFT	0.9365	0.9374	0.9393	105	70	110
		ZoomFFT on Env	0.8082	0.8109	0.8140	322	288	332

dataset, 35 observations were misclassified every time. Compared to the minimum and the maximum number of misclassified observations over these trials (37 and 42, respectively), the number of consistently misclassified observations is quite significant. This brings up a hypothesis that the application of interpretability-enhancing preprocessing makes a portion of the data impossible to classify correctly. This phenomenon seems to originate in the fact that the envelope branch, specifically the HT, is making the signals more interpretable to humans by removing some (ostensibly) irrelevant features. Nevertheless, while the removed features are irrelevant to human practitioners, they can likely be helpful to machine learning models; by their removal, a noticeable decrease in the classification performance of the models is registered.

Next, we check whether this phenomenon is independent of the first cut-off frequency of the bandpass filter since this is the most important hyperparameter of the interpretable preprocessing branch. Frequency components before the first cut-off frequency are likely to get their magnitude reduced significantly; therefore, we find this value crucial for maintaining information. We study the effect of its variation on the classification accuracy and the number of misclassified observations. In Table 3, minimum and maximum classification accuracies of MLP and the number of misclassified observations over 5 trials of the experiments for different cut-off frequencies are provided. Across these results, no difference in overall performance is seen. Although we are likely to have around 35 observations constantly misclassified for any given

Table 3: Average classification accuracy, and number of repeatably misclassified observations, on CWRU using MLP, over 5 trials ("A", "M" and "C" stand for *accuracy*, *number of misclassified* and *constantly misclassified*, respectively)

Metric	First Cut-off Frequency (Hz)														
	2.5			10			20			30			40		
A	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
	97.60	97.69	97.88	97.37	97.40	97.42	97.31	97.40	97.48	96.97	97.17	97.31	96.97	97.22	97.42
M	Min	Max	C	Min	Max	C	Min	Max	C	Min	Max	C	Min	Max	C
	37	42	35	45	46	36	44	47	37	47	53	38	45	53	36

value of the first cut-off frequency, it is to be noted that only 25 observations were *never* correctly classified across all the different frequency values.

Finally, in Table 4, the number of each combination of ground-truth and misclassified labels – of the 25 constantly misclassified observations, no matter what is the first cut-off frequency – is summarized. According to this Table, the ball problem is always either the ground truth or misclassified label, in all of these observations. This finding is a confirmation of the previously presumed hypothesis that the application of the interpretable preprocessing branch makes a portion of the data – in this case study, a relatively limited number of ball fault observations – impossible to classify correctly. By comparing Figures 2b and 2c with Figure 2d, we can see that – unlike inner race and outer race faults – the envelope preprocessing branch is not successful in revealing expected bearing fault characteristic frequency components. We believe that, alongside the missing dominant peak at the fault characteristic frequency components, the low-frequency peaks at the right subplot of Figure 2b are the reasons why ball fault signals are often misclassified.

Table 4: Types of misclassifications that occur in the envelope branch consistently, i.e., regardless of the cut-off frequency.

Ground-truth Label	Misclassified as	Count
Outer-Race Fault	Ball Problem	12
Ball Problem	Outer-Race Fault	12
Ball Problem	Normal	1

5.1 Application of SHAP to Explain Classifiers

SHAP is an explanation method originated from game theory literature [10], concerned with the calculation of an additive feature importance score [1]. The importance score of each feature is assessed by the comparison of the model performance when including and excluding the desired feature in different coalitions, computed as the weighted average of all possible differences [17].

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)], \tag{4}$$

where F is the set of all features, $f_{S\cup\{i\}}$ is the model trained with an arbitrary feature, and f_S is the model trained without that feature.

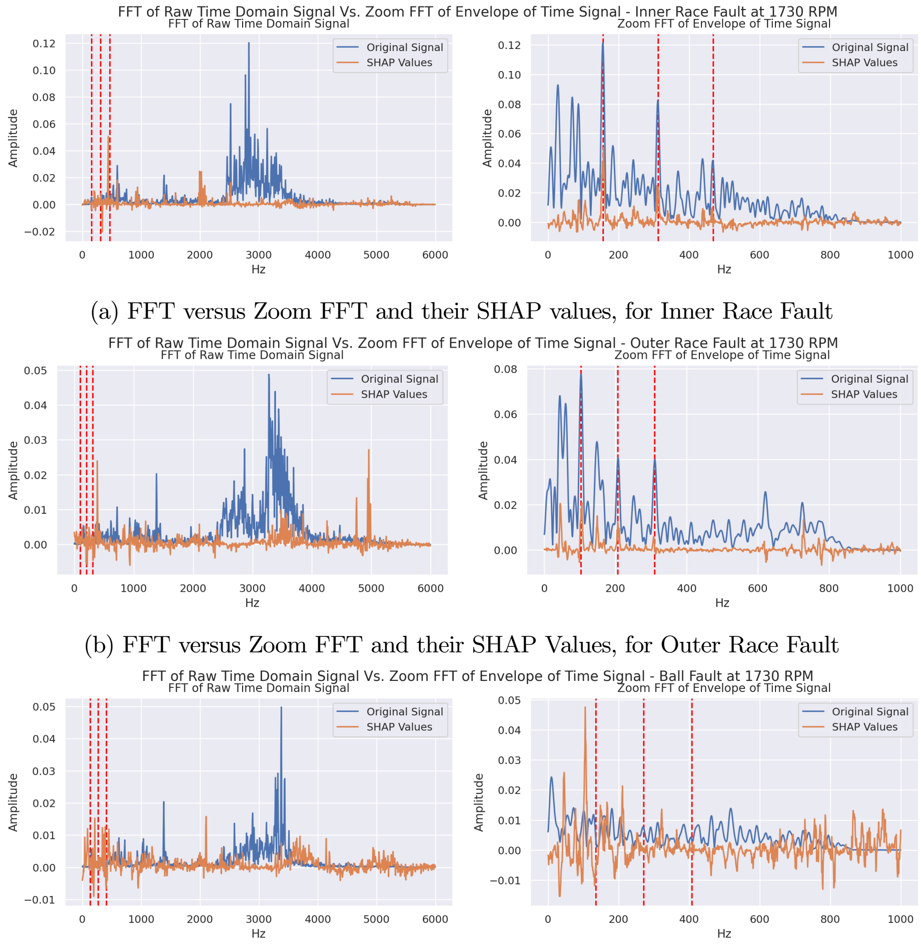
We employ SHAP (as implemented by [17]) to estimate the importance of every frequency component towards each prediction. In Figure 3, three instances of frequency domain signals – each exemplifying a fault class – from both preprocessing branches (the conventional branch on the left and the interpretable one on the right) are visualized. The input data is shown in blue, and the corresponding SHAP explanations are in orange. The red dashed lines are the first, second, and third harmonics of the fault characteristic frequency components, according to Table 1.

The perfect alignment of peaks from both original signals and SHAP values at physically expected frequencies on the right-hand subplots of Figures 3a and 3b shows that explanations from the envelope preprocessing branch match the expected physical patterns very well; the lack of the same on the left-hand subplots indicates that the opposite is true for the conventional, or raw, branch. Besides, the comparison of Figure 3c with Figures 3a and 3b shows that the agreement between explanations and the physically expected patterns varies with the type of fault. In other words, the interpretable processing branch is not capable of dealing with all the classes. While the explanations for inner and outer race faults are as expected, the ball faults are not. This can be seen in the right-hand subplot in Figure 2d, where in contrast with inner race and outer race faults, no dominant peak can be observed for the ball fault. Moreover, low-frequency peaks are likely to make this bearing fault detection harder.

Moreover, while the model utilizing the conventional preprocessing branch is likely to perform perfectly, its explanations (left-hand plots visualized in Figure 3) show no meaningful alignment with the physically expected patterns. This lack of agreement with the physics knowledge is the disadvantage of this model in comparison with its interpretable counterpart and will likely make it less trustworthy.

6 Conclusions

In this study, we evaluated how the classification accuracy of bearing fault detection changes depending on including or excluding a counter-modulation technique. We ran experiments over two datasets and used four classification algorithms. Results show that while the demodulated pipeline offers higher interpretability, aligning better with the underlying physical phenomena, its classification performance is decreased noticeably. Therefore, we believe an inherent interpretability versus performance trade-off exists from the data-centric (alternatively to be called representation, feature extraction, or preprocessing) perspective. With complex enough problems, making the data representation interpretable involves simplifications that remove information – information that would otherwise be possible for machine learning algorithms to exploit. The effect is consistent for variations in the first cut-off frequency of the interpretable preprocessing branch, different datasets, and classification algorithms.



(a) FFT versus Zoom FFT and their SHAP values, for Inner Race Fault

(b) FFT versus Zoom FFT and their SHAP Values, for Outer Race Fault

(c) FFT versus Zoom FFT and Their SHAP Values, for Ball Fault

Fig. 3: Examples of each fault from both branches, with SHAP values

Our supplementary analysis shows that applying the envelope preprocessing branch affects a relatively minor portion of the data. We believe this is due to removing the features irrelevant to human analysts and simultaneously useful for AI models. The next step in pursuing this study is to understand the adversarial mechanism responsible for this decrease, hopefully leading to the discovery of transformations with a better balance between the two aspects.

Furthermore, since some of the misclassified samples differed between experiments with different first cut-off frequencies, this hyperparameter can be considered a factor in generating diverse datasets. It may be, therefore, possible to improve fault classification accuracy by using an ensemble of different datasets produced by varying the first cut-off frequencies.

Current results demonstrate the idea in a single domain. It is interesting to extend this research and explore this data-centric interpretability versus performance trade-off in other fields where well-understood interpretable transformations exist, such as computer vision or speech recognition.

Acknowledgements: This work was partially supported by Vinnova and by CHIST-ERA grant CHIST-ERA-19-XAI-012 from Swedish Research Council.

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion **58**, 82–115 (2020)
2. Bechhoefer, E.: A quick introduction to bearing envelope analysis. Green Power Monit. Syst (2016)
3. Brito, L.C., Susto, G.A., Brito, J.N., Duarte, M.A.: An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mechanical Systems and Signal Processing **163**, 108105 (2022)
4. Chen, H.Y., Lee, C.H.: Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis. IEEE Access **8**, 134246–134256 (2020)
5. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: A survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). pp. 0210–0215. IEEE (2018)
6. Fan, Y., Hamid, S., Nowaczyk, S.: Incorporating physics-based models into data-driven approaches for air leak detection in city buses. In: ECML PKDD 2022 Workshops (2022)
7. Feldman, M.: Hilbert transforms. In: Braun, S. (ed.) Encyclopedia of Vibration, pp. 642–648. Elsevier, Oxford (2001)
8. Feldman, M.: Hilbert transform in vibration analysis. Mechanical systems and signal processing **25**(3), 735–802 (2011)
9. Han, D., Liang, K., Shi, P.: Intelligent fault diagnosis of rotating machinery based on deep learning with feature selection. Journal of Low Frequency Noise, Vibration and Active Control **39**(4), 939–953 (2020)
10. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI methods-a brief overview. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. pp. 13–38. Springer (2022)
11. Lee, D.H., Hong, C., Jeong, W.B., Ahn, S.: Time–frequency envelope analysis for fault detection of rotating machinery signals with impulsive noise. Applied Sciences **11**(12), 5373 (2021)
12. Lee, J.S., Yoon, T.M., Lee, K.B.: Bearing fault detection of ipmsms using zoom FFT. Journal of Electrical Engineering and Technology **11**(5), 1235–1241 (2016)
13. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K.: Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing **138**, 106587 (2020)
14. Lessmeier, C., Kimotho, J.K., Zimmer, D., Sextro, W.: Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In: PHM Society European Conference. vol. 3 (2016)

15. Li, C., Zhang, W., Peng, G., Liu, S.: Bearing fault diagnosis using fully-connected winner-take-all autoencoder. IEEE Access **6**, 6103–6115 (2017)
16. Liu, Y.: Fault diagnosis based on SWPT and Hilbert transform. Procedia Engineering **15**, 3881–3885 (2011)
17. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
18. Meng, Z., Zhan, X., Li, J., Pan, Z.: An enhancement denoising autoencoder for rolling bearing fault diagnosis. Measurement **130**, 448–454 (2018)
19. Mey, O., Neufeld, D.: Explainable AI algorithms for vibration data-based fault detection: Use case-adapted methods and critical evaluation. arXiv preprint arXiv:2207.10732 (2022)
20. Rajabi, S., Azari, M.S., Santini, S., Flammini, F.: Fault diagnosis in industrial rotating equipment based on permutation entropy, signal processing and multi-output neuro-fuzzy classifier. Expert Systems with Applications **206** (2022)
21. Rudin, C., Radin, J.: Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. Harvard Data Science Review **1**(2) (nov 22 2019), https://hdsr.mitpress.mit.edu/pub/f9kuryi8
22. Wang, N., Liu, X.: Bearing fault diagnosis method based on Hilbert envelope demodulation analysis. In: IOP Conference Series: Materials Science and Engineering. vol. 436, p. 012009. IOP Publishing (2018)
23. Xia, M., Li, T., Liu, L., Xu, L., de Silva, C.W.: Intelligent fault diagnosis approach with unsupervised feature learning by stacked denoising autoencoder. IET Science, Measurement & Technology **11**(6), 687–695 (2017)