Engineering Applications of Artificial Intelligence 136 (2024) 109046

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Relation between fault characteristic frequencies and local interpretability shapley additive explanations for continuous machine health monitoring

Tongtong Yan [a], Xueqi Xing [a], Tangbin Xia [b,c], Dong Wang [b,c,*]

[a] Department of Mechanical and Materials Engineering, University of Western Ontario, London, Ontario, Canada
[b] The State Key Laboratory of Mechanical Systems and Vibration, Shanghai Jiao Tong University, Shanghai, 200240, PR China
[c] Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, PR China

ARTICLE INFO

Keywords:
Shapley additive explanations
Fault characteristic frequencies
Machine health monitoring
Health indicator
Signal filtering method

ABSTRACT

Recently, the Shapley additive explanations models have been extensively studied to enhance explainability of artificial intelligence algorithms, while most of them simply use Shapley additive explanations to rank or measure the importance of different features. In this study, a novel methodology that studies the relation between fault characteristic frequencies and Shapley values generated by local interpretability Shapley additive explanations for machine health monitoring is proposed. Firstly, a simulation model is introduced to generate vibration signals at different health conditions and their spectral amplitudes transformed from Fourier transform are used to investigate the relationship between fault characteristic frequencies and local interpretability Shapley values. It is interestingly found that Shapley values can be used to locate fault characteristic frequencies. Moreover, most of them have negative values in a normal stage and have positive values in an abnormal stage. Based on this finding and Shapley additive explanations, a health indicator construction methodology is proposed to continuously monitor incipient machine faults. Subsequently, an automatic signal filtering method is proposed to remove and eliminate burrs and noise in Shapley values so that fault characteristic frequencies can be clearly revealed by Shapley values for physical fault diagnosis. Two run-to-failure cases are conducted to demonstrate the effectiveness of the proposed methodology and then the superiority of this study is demonstrated by comparing with existing methods for health indicator construction and fault diagnosis, including sparsity parameters, Hjorth parameters, and fast Kurtogram. Comparison results show that the proposed health indicator is more sensitive to the time of incipient fault initiation and interpretable fault diagnosis based on Shapley values has a robust performance. This study first sheds a light on the relationship between fault characteristic frequencies and Shapley values under the scenario of continuous machine health monitoring and seamlessly guides applicants to realize Shapley additive explanations based incipient fault detection and diagnosis.

1. Introduction

Continuous machine health monitoring is a long-term monitoring task to realize timely incipient fault detection and subsequently diagnose incipient fault types (Zhou et al., 2022). Without external disturbances, a machine initially runs in a normal stage for a long time. As the performance of a machine gradually degrades over time, it will enter an incipient fault stage and slight faults will occur (Dibaj et al., 2020). In engineering, this is a best time to implement some necessary maintenance so that machine life can be greatly prolonged to avoid catastrophic failure. Not only the time of incipient faults is discovered but also its exact fault type needs to be confirmed for timely carrying out targeted and accurate repairs (Yu et al., 2024). Therefore, machine health monitoring is a core technology to realize condition-based maintenance (CBM) (Jardine et al., 2006).

In the domain of machine health monitoring, eXplainable artificial intelligence (XAI) is an emerging technology. Since the black-box property of AI algorithms hinders their reliable applications in engineering and it is difficult for users to trust and understand how these models work, many studies were proposed to enhance the interpretability of their models for machine health monitoring (An et al., 2023; Yan et al., 2022; Jia et al., 2018; Jin et al., 2023; Feng et al., 2023).

* Corresponding author. The State Key Laboratory of Mechanical Systems and Vibration, Shanghai Jiao Tong University, Shanghai 200240, PR China.
E-mail addresses: tongtongyan97@outlook.com (T. Yan), xxing53@uwo.ca (X. Xing), xtbxtb@sjtu.edu.cn (T. Xia), dongwang4-c@sjtu.edu.cn (D. Wang).

https://doi.org/10.1016/j.engappai.2024.109046
Received 21 October 2023; Received in revised form 25 March 2024; Accepted 22 July 2024
Available online 26 July 2024

However, most of them just focused on a specific model without generality or combined data-driven models with physical models to achieve ante-hoc or post-hoc interpretability (Lipton, 2018). Ante-hoc interpretability refers to the employment of a straightforward and understandable self-explanatory model or the incorporation of interpretability into a model structure to obtain built-in interpretability of the model itself before model training. Self-explanatory models, such as linear models and decision trees have ante-hoc interpretability. After fully learning a model, post-hoc interpretability is performed to use explanatory techniques or construct explanatory models to clarify model's working mechanism or rationale for making decisions. For example, since wavelet transform and convolutional neural network (CNN) both have convolution operation, Li et al. (2021) embedded a wavelet convolutional layer into CNN to achieve its partial ante-hoc interpretability for machine fault diagnosis. Wang et al. (2022) designed a Fourier transform, wavelet transform, and square envelope driven neural network to realize fully ante-hoc interpretability for machine health monitoring. An et al. (2023) integrated algorithm unrolling of a sparse coding model with a generative adversarial network for machine fault diagnosis to achieve ante-hoc interpretability. Most ante-hoc interpretability techniques are model-specific and they incorporate physical elements with the structures of AI algorithms. As for post-hoc interpretability, Yan et al. (2022) proposed a Fisher criterion based optimization model to identify informative frequency bands based on weight variables for machine health monitoring. Li et al. (2022) utilized attention mechanisms and transformer networks to obtain attention heat maps for useful sample identification. More works for machine health monitoring based on XAI can be seen in (Sadoughi and Hu, 2019; Shen et al., 2021; Yang et al., 2021).

In addition, some popular and latest methods or tools have been widely used to explain the prediction results of AI models, such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), and Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017). In the realm of mechanical health monitoring, these models play a pivotal role in determining the relative significance of input features in relation to fault detection or diagnostic outputs generated by AI models, thus providing crucial insights into a prediction process. The Grad-CAM is a technique to interpret CNN decisions, which uses gradient information to identify which regions in an image contribute the most to predicting specific categories. Through this method, a heatmap can be generated to display which parts of the image are most important for the model's decision-making. Therefore, time-frequency spectra of vibration signals under different conditions can be combined with Grad-CAM to explain the diagnostic outputs of CNNs. Li et al. (2023) proposed a multilayer Grad-CAM to obtain activation maps under various resolutions for interpretable bearing fault diagnosis. To eliminate prior and expert knowledge, Yoo et al. (Yoo and Jeong, 2022) presented a vibration analysis procedure that utilizes a CNN model in conjunction with Grad-CAM to approximately pinpoint informative frequency bands within spectrogram images. Nevertheless, Grad-CAM aided XAI for machine fault detection and diagnosis mainly focused on time-frequency spectrograms and provided an approximate location in the time-frequency domain. The LIME belongs to a permutation importance method and it quantifies a feature's influence on a model by randomly shuffling its values and subsequently monitoring alterations in accuracy. However, LIME methods are sensitive to sample selection. SHAP model was proposed to realize the interpretability of black-box models. The basic idea of the SHAP model is to assess the relative importance of input variables using Shapley values from game theory. Since the SHAP model only requires knowledge of a black-box model's output for an input sample's neighboring instances, it is model-agnostic. Moreover, each observed feature of a sample receives its own SHAP value or Shapley value, which can realize local interpretability. Since it was first proposed, this approach has drawn a lot of interests and has been widely used in numerous fields, such as metallogenic prediction (Fan et al., 2023), alloy design (Liu et al., 2023), and manufacturing processes (Rocha et al., 2022). In terms of machine health monitoring, Souza et al. (Hoffmann Souza et al., 2023) combined Autoencoder (AE) with the SHAP model to analyze the behaviors of different features and their relevance to failures. Xu et al. (2022) used the SHAP model to select optimal feature space for water pipe leakage monitoring under different scenarios and algorithms. Jakubowski et al. (2021) demonstrated the application of a variational autoencoder to track wear on rolls in a hot strip machine and used the SHAP model to understand the contributions of different measurements. More similar ideas have been reported in (Park et al., 2022; Decker et al., 2023). Although numerous attempts have been made to explore the use of the SHAP model for machine health monitoring, the majority of them merely employed SHAP values to rank or measure the significance of various features/measurements without taking further consideration. For example, how the distribution and characteristics of SHAP values change under different health conditions should be explained. In this way, SHAP values can be well utilized to guide users to develop some interpretable methodology for machine health monitoring. Furthermore, the study of SHAP values for health indicator (HI) construction for incipient fault detection during continuous health monitoring is seldom reported.

To fill in these gaps, in this study, a novel methodology is proposed by studying the relation between fault characteristic frequencies and local interpretability SHAP values for continuous health monitoring. The main contributions of this study are illustrated as follows. Firstly, the relations between fault characteristic frequencies and local interpretability SHAP values are thoroughly investigated based on a simulation model that can generate vibration signals under different health conditions. An interesting conclusion can be drawn that peaked SHAP values can be used to locate fault characteristic frequencies. Herein, negative SHAP values concerning fault characteristic frequencies can be observed in a normal stage while their positive values can be observed in an abnormal stage. This finding can directly facilitate establishing some rules for machine health monitoring. Secondly, based on this finding, a HI based on SHAP values is proposed to monitor incipient machine faults by threshold setting. This study first reveals a sight to utilize SHAP values for HI development. Finally, a burr filtering and noise removal methodology of SHAP values is proposed to explicitly show fault characteristic frequency and its harmonics for physical fault diagnosis after the time of incipient fault initiation.

The structure of this paper is organized as follows. In section 2, a proposition of SHAP values and fault characteristic frequencies is first proposed and it is experimentally investigated based on a simulation model. Afterward, a novel HI construction and signal filtering methodology based on SHAP values is proposed for continuous machine health monitoring. Finally, the whole flowchart of the proposed methodology is summarized. Section 3 demonstrates applications of the proposed methodology on two run-to-failure cases. Conclusions are drawn in a final section.

2. Material and methods

2.1. Investigation on the relation between fault characteristic frequencies and local interpretability SHAP values based on a simulation study

In this section, a simulation study that can generate vibration signals under different health statuses is used to explore the relation between fault characteristic frequencies and local interpretability SHAP values. Firstly, the fundamental theory of the SHAP model is presented. Although the SHAP model can be applied to any black-box model with local accuracy, consistency, and missingness properties, its computational load is extremely large. To reduce computational complexity, a variant of the SHAP model called TreeSHAP is used and it is a fast computation of the SHAP model to calculate SHAP values based on tree-based AI models, such as random forests and gradient-boosted trees.

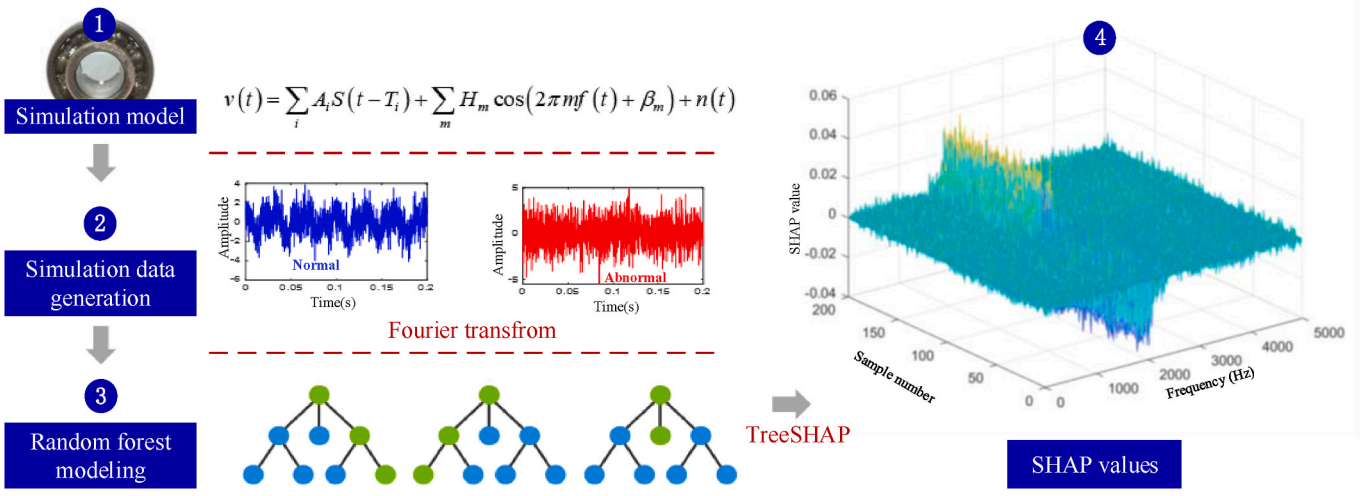

Fig. 1. Flowchart of investigating the relation between fault characteristic frequencies and local interpretability SHAP values based on a simulation study.

Therefore, in this study, a random forest is used as a base model to acquire SHAP values. Afterward, normal and abnormal vibration signals are generated based on a simulation model and a binary-classification random forest model is trained by their spectral amplitudes transformed from Fourier transform (FT). Once the random forest model is trained and established, the SHAP values of spectral amplitudes of each vibration signal can be accordingly calculated. Finally, the relation between fault characteristic frequencies and local interpretability SHAP values is explored and analyzed. A whole idea to investigate the relation between fault characteristic frequencies and local interpretability SHAP values based on a simulation study is summarized in Fig. 1.

2.1.1. SHAP and TreeSHAP models

The majority of AI-based models now in use are black boxes, making it challenging to interpret and understand model predictions and their relationships with input features. In a black-box model, it is difficult to figure out the underlying rules and mechanisms behind model prediction. Although AI-based models have great potential in extracting discriminating features and establishing nonlinear models in engineering, their black box property cannot guarantee their reliability and further hinder their applications. This is because users cannot see decision processes made by models and understand working mechanisms based on their experience and knowledge. To solve this dilemma, Lundberg and Lee (2017) proposed a milestone technology named the SHAP method to explain any box-black model and it is model-agnostic to interpret every output made by the AI model. The SHAP technique was created by combining concepts from game theory with local interpretations. The main idea of the SHAP is to calculate a feature's marginal contribution when it is added to a model and this concept originates from Shapley values in cooperative game theory. Given all feature combinations, the SHAP values stem from an additive feature attribution method as follows.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j, \tag{1}$$

where M means the number of input features and it is equal to the number of spectral amplitudes of each sample transformed from FT in our study. g represents an explanatory model. ϕ_0 is a constant term and $\phi_j \in R$ is the feature attribution of j_{th} feature. z'_j is a simplified feature and it can only take from 0 or 1 as follows.

$$z'_j = \begin{cases} 0, \text{the } j_{th} \text{ feature is present} \\ 1, \text{the } j_{th} \text{ feature is not present} \end{cases}, \tag{2}$$

In equation (1), all ϕ_j that are added together roughly match the output of an initial model. Moreover, the additive feature attribution method in equation (1) has some appealing characteristics, such as local accuracy, missingness, and consistency (Lipovetsky, 2022). Herein, the meaning of local accuracy is that the output of the model we are trying to understand is equal to the total of the feature attributions. Missingness refers to a feature that has already vanished, such as $z'_j = 0$ and its attribution is equal to zero. Consistency states that when an observed feature has a greater impact on a model, altering a model will never result in a reduction in the attribution given to that feature. Therefore, given a sample $\mathbf{x}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{iM}]$ and an original model $f(x)$, its SHAP values satisfy the equation as follows.

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{iM}), \tag{3}$$

where the predicted value of $\mathbf{x}_i$ is y_i. The mean of the predicted values across all samples is y_{base} and it serves as a baseline for the model. $f(x_{ij})$ is the SHAP value of x_{ij}. It can be observed from equation (3) that subtracting the average predicted value from the predicted value is equal to the contributions of all m features given a linear model. Although the SHAP model is theoretically strongly guaranteed, calculating SHAP values for features is computationally expensive and generally requires exponential time.

Subsequently, a fast computation of SHAP values called the TreeSHAP model was accordingly proposed by Lundberg et al. (2018) and it is a variation of the classical SHAP model. The original model of the TreeSHAP should be tree-based AI models, such as random forests and gradient-boosted trees. Lundberg et al. (2018) pointed out that the TreeSHAP is a polynomial time algorithm to compute optimal explanations grounded in game theory principles. Besides, based on three medical datasets, it was shown that Tree-based models can exhibit higher accuracy than neural networks while maintaining greater interpretability compared to linear models. Therefore, the TreeSHAP is a main focus of this study to obtain SHAP values of features and it is used to investigate the relationship between fault characteristic frequencies and local interpretability SHAP values. The brief usage processes of TreeSHAP are provided as follows. Firstly, a tree model needs to be trained based on a dataset to correlate input features and output results. After the tree model training is completed, use the TreeSHAP algorithm to calculate the contribution of each feature to prediction results, namely SHAP values. The process of calculating SHAP values typically involves recursively traversing all possible paths of the tree model and calculating the weighted contribution of each feature on different paths. Specifically, the TreeSHAP algorithm first recursively traverses each node of the tree model, tracking the flow of different subsets based on

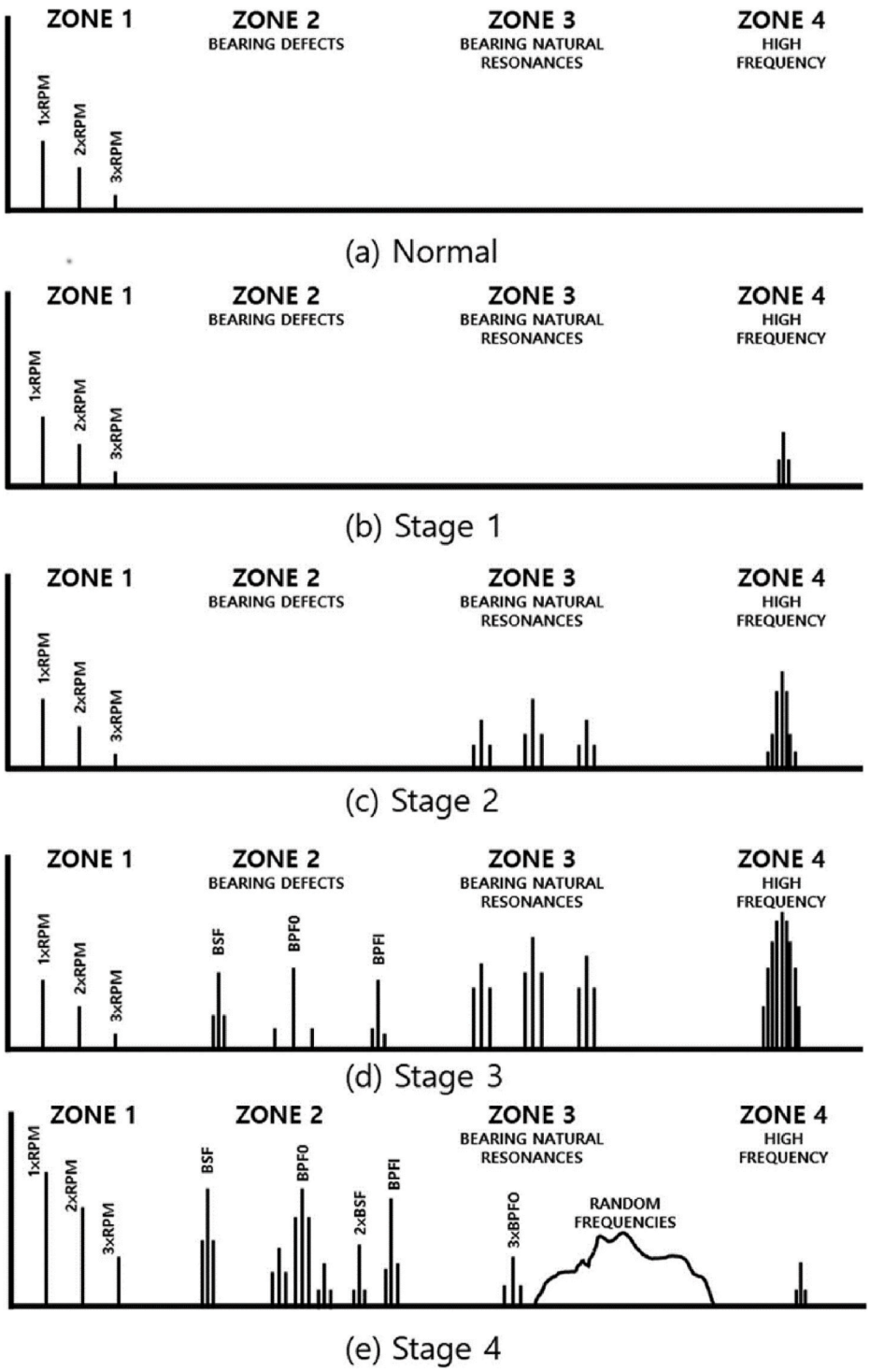

Fig. 2. Evolution of spectral amplitudes in a whole frequency zone during long-term degradation processes (Park et al., 2021).

the segmentation conditions of the features. For each leaf node, TreeSHAP calculates the contribution of the features on all possible paths to reach that leaf node. Subsequently, TreeSHAP takes into account the weight of the features on each path based on the coverage of the features and segmentation conditions. By weighting contributions, the algorithm can quantify the relative importance of each feature to prediction results. Assume that T represents the number of trees, D stands for the maximum depth of any given tree, and L signifies the number of leaves. The worst-case complexity of the TreeSHAP is $O(TLD^2)$. The pseudocode of the TreeSHAP is summarized in Algorithm 2 of reference (Lundberg et al., 2018) and some specialized libraries such as Python's shap library can be used to implement the calculation of TreeSHAP. In this study, random forest (Breiman, 2001) is chosen as a base model and it is an ensemble tree model composed of multiple decision trees. By introducing randomness, the random forest has a good noise resistance property and quick training times to enable model parallelization. The training process of a random forest classifier requires multiple stages. Initially, separate decision trees are constructed, each customized for a unique subset of training data. These subsets are created by randomly sampling and replacing an original dataset to ensure the diversity

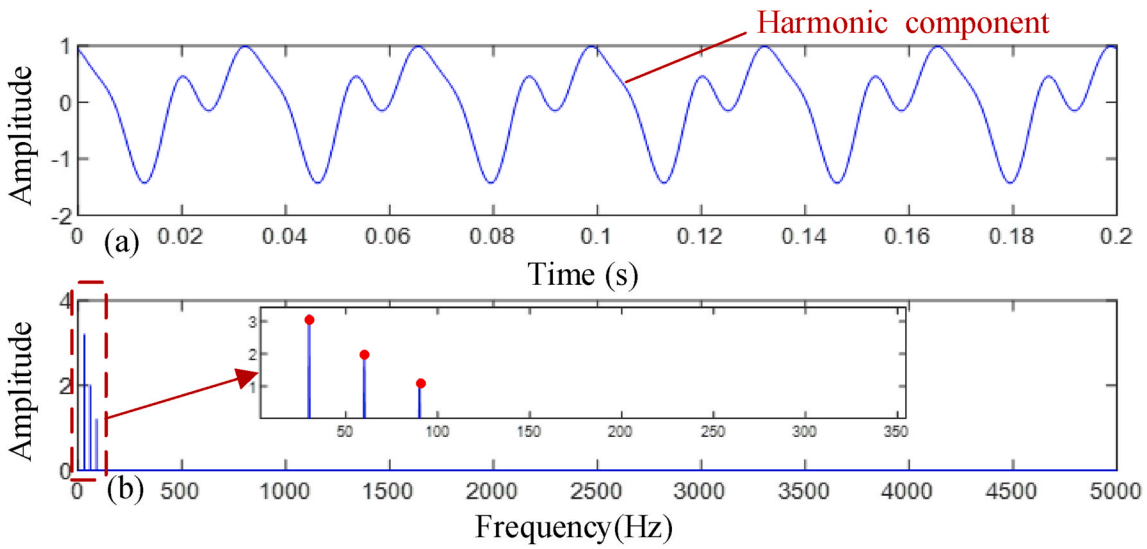

Fig. 3. Simulated synthesized harmonic components and their corresponding spectral amplitudes: (a) harmonic components in the time domain; (b) harmonic components in the frequency domain.

between trees. In addition, random feature selection is performed at each node of the tree to guide a segmentation process. Finally, the classification results are determined through a voting mechanism, where predictions for all trees in the forest are aggregated to produce a final output. This integration method allows random forest classifiers to utilize the collective intelligence of multiple trees, thereby improving classification accuracy and reducing overfitting risks. In this study, when training a random forest model, the labels of normal spectral amplitudes are set to 1 and the labels of abnormal spectral amplitudes are set to -1. Before starting training, it is necessary to determine some parameters, such as the number of trees in the random forest, the depth of each tree, the number of features used by each node, and termination conditions. Some key parameters of random forest in training processes and model establishment are given as follows. The number of trees is set to 10 and the depth of each tree is not limited. A Gini index is used to measure the degree of impurity in the random forest model. Once the random forest model is trained based on normal and abnormal spectral amplitudes, the TreeSHAP algorithm is applied to calculate the SHAP values of each spectral line for each sample.

2.1.2. Relation between fault characteristic frequencies and local interpretability SHAP values based on the TreeSHAP model and a simulation model

To investigate the correlation between fault characteristic frequencies and SHAP values for continuous machine health monitoring, we initially analyze the evolution and alterations of these frequencies during long-term degradation processes. Illustrated in Fig. 2, the distribution of spectral amplitudes in the frequency domain closely correlates with machine degradation processes. The entire frequency range comprises four zones: low-frequency, defect frequency, natural resonance frequency, and high-frequency. Fig. 2(a)–(e) reveal that shaft rotation frequencies and their harmonics primarily occur and always exist in the low-frequency zone, regardless of machine conditions. As the machine transitions to abnormal status and initial faults emerge, cyclic fault frequencies contaminated by noise arise in the high-frequency zone, as seen in Fig. 2(b). A limited sampling frequency makes acquiring amplitudes in this zone challenging. Instead, it is more effective to detect incipient faults in the natural resonance frequency zone, where fault frequencies with equal spacing appear, as in Fig. 2(c). As faults become worse, more characteristic frequencies and harmonics emerge in the defect frequency zone, as shown in Fig. 2(d)–(e). In summary, fault frequencies and their harmonics in the defect and natural resonance frequency zones become more prominent as faults develop and they are crucial for detecting and diagnosing incipient faults.

Based on the above analysis, a *proposition* about the relation between fault characteristic frequencies and local interpretability SHAP values is proposed as follows.

Proposition 1: Given a binary classification model based on random forest, obtained SHAP values of spectral amplitudes can be used to locate fault characteristic frequencies and they have negative values in a normal stage and positive values in an abnormal stage.

Proposition 1 is experimentally explored by using a simulation model that can generate vibration signals at different health conditions. A fault vibration simulation model can be formulated as follows (Zhao et al., 2013).

$$v(t) = \sum_i A_i S(t - T_i) + \sum_m H_m \cos(2\pi m f(t) + \beta_m) + n(t), \tag{4}$$

and,

$$S_i(t) = \mathrm{e}^{-\zeta t} \sin(2\pi f_r t), \tag{5}$$

where $v(t)$ is the simulated fault vibration signal and it contains three components. The first component imitates repetitive transients $\sum_i A_i S(t - T_i)$ in the time domain caused by faults. It can be observed from equation (5) that the fault impulses are modeled by an exponentially decaying sinusoid. Such fault signatures of vibration signals usually appear when local faults occur in rotating machinery, such as bearings. For example, when a localized bearing fault occurs, rollers impact the faulty area, resulting in the generation of impulsive transients. A_i represents the intensity of the i_{th} pulse and it is the pulse amplitude. T_i is the time of i_{th} pulse occurrence. ζ means the damping ratio of fault impulses and f_r is the resonance frequency. The second component imitates synthesized deterministic frequency components $\sum_m H_m \cos(2\pi m f(t) + \beta_m)$ in the time domain and they are composed of the fundamental frequency of the shaft and its corresponding harmonics. Herein, m is the number of harmonic components and H_m represents the amplitude of the m_{th} harmonic component. $f(t)$ and β_m respectively denote the instantaneous rotating frequency and the initial phase of the m_{th} harmonic component. The last term $n(t)$ indicates measurement noise and it can be set by random Gaussian noise. Therefore, normal vibration signals can be generated by ignoring the fault-inducing components as follows.

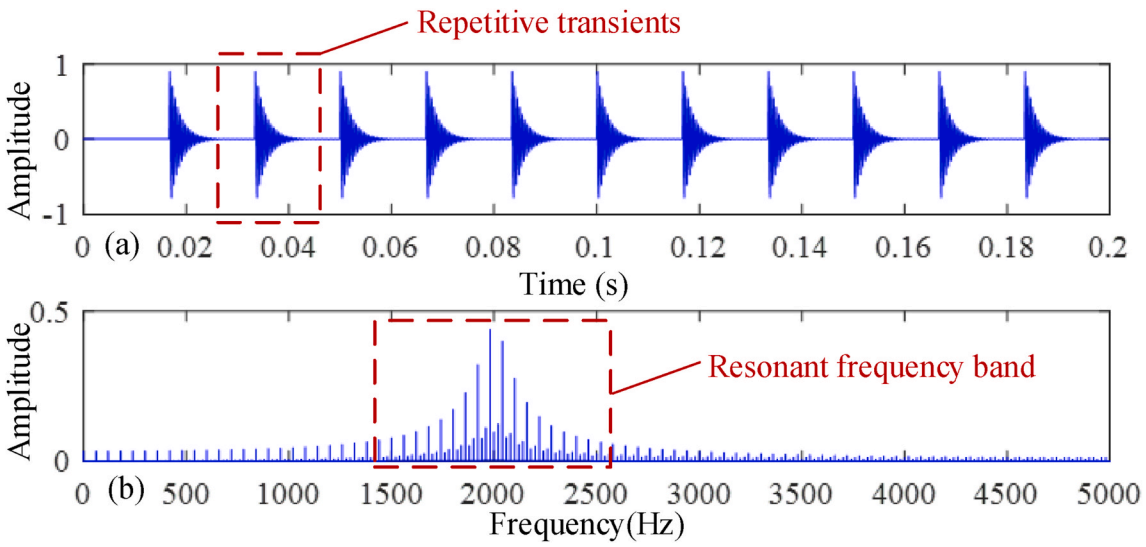

Fig. 4. Simulated repetitive fault transients and their corresponding spectral amplitudes: (a) repetitive fault transients in the time domain; (b) repetitive fault transients in the frequency domain.

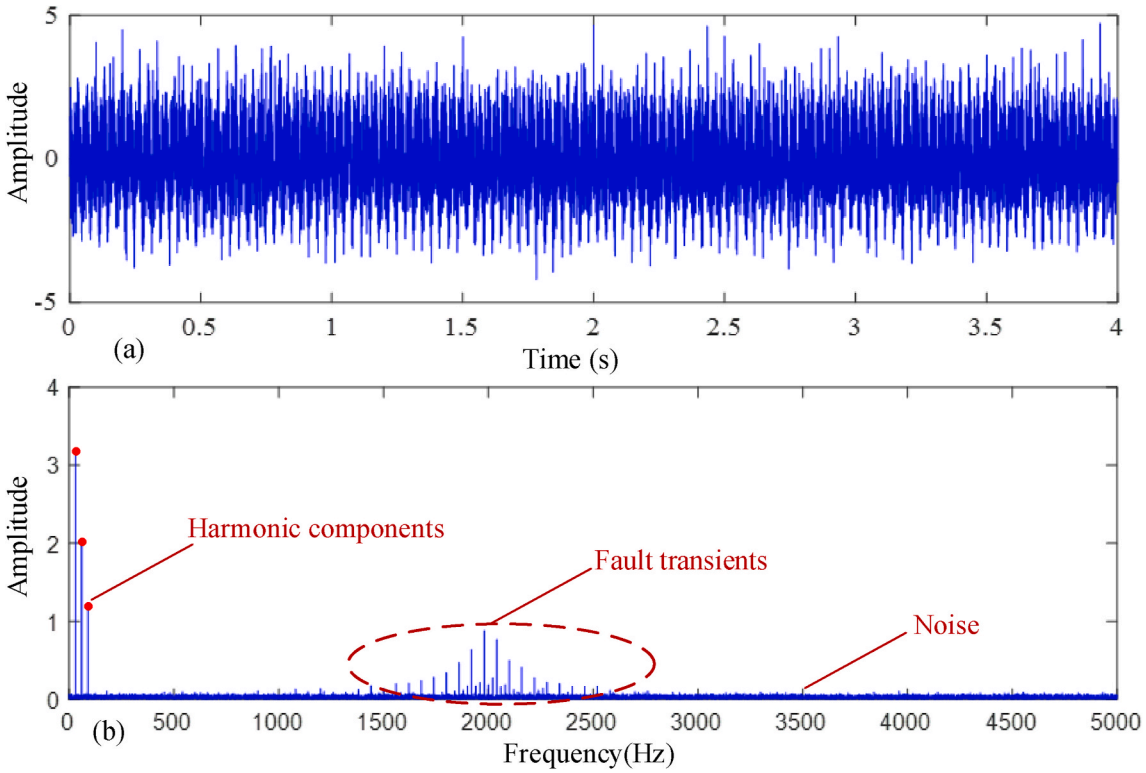

Fig. 5. Simulated synthesized signals of harmonic components, repetitive transients, and random Gaussian noise and their corresponding spectral amplitudes: (a) synthesized signals in the time domain; (b) synthesized signals in the frequency domain.

$$v^n(t) = \sum_m H_m \cos(2\pi m f(t) + \beta_m) + n(t), \tag{6}$$

In our study, three harmonic signals with instantaneous frequencies of 30, 60, and 90 are synthesized to generate the fundamental frequency of the shaft and its corresponding two harmonics, and their amplitude intensities are set as 0.8, 0.5, and 0.3. A simulated harmonic signal in the time domain and its corresponding spectral amplitudes in the frequency domain, obtained through FT, are illustrated in Fig. 3. Three distinct spectral lines are visible in the frequency domain. Nevertheless, both normal and faulty machine states are susceptible to noise interference. Consequently, simulated vibration samples for the normal operating stage are created by combining deterministic frequency components with random Gaussian noise.

For fault-inducing components, A_i, T_i, ζ and f_r are respectively set as 1, $\frac{1}{59}$, 500, and 2000. Under these parameters, Fig. 4 depicts the repetitive fault transients in the time domain along with their spectral amplitudes. An ideal fault signal exhibits distinct signatures in both domains. In the frequency domain, the energy primarily concentrates around the natural resonance frequency. Specifically, the repetitive transients in the time domain manifest as parallel spectral lines,

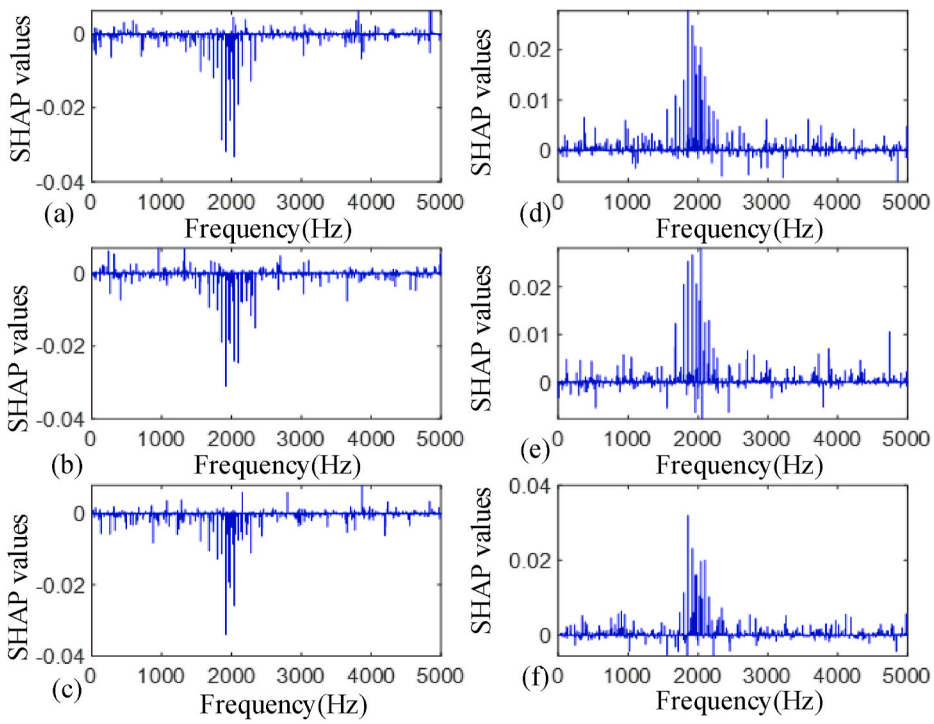

Fig. 6. Obtained SHAP values of some representative samples: (a) 1_{th} sample; (b) 50_{th} samples; (c) 100_{th} sample; (d) 101_{th} sample; (e) 150_{th} sample; (f) 200_{th} sample.

primarily situated within the resonant frequency band. These spectral lines are evenly spaced, with intervals corresponding to the fault characteristic frequencies. However, these fault signatures in both domains can easily be overwhelmed by significant noise, particularly during the early stages of fault development. The simulated fault vibration samples are created by integrating deterministic frequency components, repetitive transients, and random Gaussian noise, as illustrated in Fig. 5.

Based on the above simulation models, vibration samples at different health conditions can be generated to facilitate the investigation on the relation between fault characteristic frequencies and local interpretability SHAP values. In this study, 100 vibration samples under normal and abnormal conditions are respectively generated and their sample lengths are equal to 10000. A sampling rate is set to 10000 Hz. The frequency resolution of spectral lines is dependent on both the length of a sample utilized and the sampling frequency. In the simulation studies, 10000 vibration sampling points from each sample are transformed into spectral lines through the application of the FT. Therefore, the spectral resolution in the frequency domain is 1Hz and 5000 spectral amplitudes ranging from 1 Hz to 5000 Hz can be obtained due to symmetrical characteristics. A total of 200 vibration samples can be obtained and they are used to simulate changes from the normal to abnormal state of a machine. The occurrence time of anomaly is the 101_{th} sample. Based on the TreeSHAP model and these simulated vibration samples, the steps to study relations between fault characteristic frequencies and local interpretability SHAP values are given as follows.

Step 1: generate 100 vibration samples under normal and abnormal conditions;
Step 2: transform each vibration signal in the time domain into spectral amplitudes in the frequency domain based on FT;
Step 3: establish a random forest model for classification based on spectral amplitudes under normal and abnormal conditions;
Step 4: use the TreeSHAP algorithm to evaluate the SHAP values of the feature for each sample.

The TreeSHAP model offers local interpretability, generating SHAP values for each sample. The SHAP values of spectral amplitudes for representative samples under both normal and abnormal conditions are illustrated in Fig. 6. The SHAP values of spectral amplitudes for the 1^{th}, 50^{th} and 100^{th} samples in a normal state are displayed in Fig. 6(a)–(c), revealing a concentration on fault characteristic frequencies within the natural resonance frequency zone, and they are primarily negative. Conversely, Fig. 6(d)–(f) presents the SHAP values for the 101^{th}, 150^{th} and 200^{th} samples in an abnormal state, exhibiting an inverse pattern. This phenomenon is inherent and can be rationalized through our physical understanding. Constant deterministic frequency components and random noise offer the limited discrimination between normal and abnormal spectral amplitudes, resulting in insignificant SHAP values for spectral amplitudes. In contrast, fault characteristic frequencies in defect and natural resonance zones are crucial for fault detection. Negative SHAP values for these frequencies in a normal state indicate their negative contributions to normal classification, biasing the random forest model towards abnormal classification. Conversely, positive SHAP values in an abnormal state aid in identifying abnormal samples. Therefore, peaked SHAP values concentrate on fault characteristic frequencies in these zones, exhibiting opposite signs based on health status.

This finding is very meaningful and inspiring. Usually, signal processing based fault detection and diagnosis method requires adept knowledge while the data-driven method lacks transparency and interpretability. However, SHAP values can serve as physics-informed features to facilitate a decision process for machine fault detection and diagnosis. For example, we can first train a random forest model based on vibration samples. The distribution of SHAP values of test data can be used to identify their status and they can give a quick intuition to realize machine health monitoring. However, it can be observed from Fig. 6 that the SHAP values have many burrs that may cause interferences and it is not convenient to directly use SHAP values for continuous machine health monitoring to confirm the precise time of incipient fault initiation. Based on the useful conclusion about the relation between fault characteristic frequencies and local interpretability SHAP values, a HI construction methodology is proposed based on SHAP values for incipient fault detection during continuous machine health monitoring. Moreover, an automatic signal filtering method of SHAP values is also introduced to remove the burrs so that fault characteristic frequencies

can be explicitly indicated and revealed by SHAP values for fault diagnosis.

2.2. Proposed HI construction and signal filtering method based on SHAP values for continuous machine health monitoring

Based on the useful finding, a novel HI construction and signal filtering method based on SHAP values is proposed for continuous machine health monitoring. Herein, a HI is continuously generated based on time-varying SHAP values to discover the time of incipient fault initiation. Once the time of an incipient fault is determined, an automatic signal filtering method is proposed to eliminate burrs in time-varying SHAP values for fault diagnosis.

2.2.1. HI construction methodology based on SHAP values for incipient fault detection

It can be seen from Fig. 6 that SHAP values show distinguishable signatures under normal and abnormal conditions, but they just provide an intuitive and qualitative perspective to identify machine health conditions. To quantitatively monitor incipient machine faults, a HI construction methodology is proposed based on SHAP values as follows.

$$HI(j) = \sum_{i=1}^{M} F_{j,i} Shap_{j,i}, j = 1, 2, \cdots, N, \tag{7}$$

where $HI(j)$ represents the HI at file number j. N is the total number of data files collected from the run-to-failure process of a machine. These data files are consecutively gathered and their timelines provide important information about the degradation processes of a machine from a normal stage to an abnormal stage and finally out of its function. M is the number of features of each sample and it means the number of spectral lines contained in each file number transformed from FT in this study. This is because spectral amplitudes at specific frequencies are directly regarded as informative features and they are explained by using SHAP values. $F_{j,i}$ represents the amplitude of the i_{th} spectral line at the j_{th} file number. $Shap_{j,i}$ is the corresponding SHAP value of $F_{j,i}$. In equation (7), SHAP values can be understood as time-varying weights of spectral amplitudes for HI construction. Herein, the amplitudes at a specific frequency among different data files represent the changing processes of this feature during the degradation processes. SHAP values can measure their contributions that vary with the changes in machine health conditions. Therefore, SHAP values are regarded as time-varying weights changed with N time steps. The definition of HI is inspired by the finding in section 2.1. It is found that the main distribution of SHAP values has an opposite pattern and sign at normal and abnormal conditions. For spectral lines in the frequency domain, their amplitudes are defined as the modulus of complex numbers and they are always positive. Therefore, the HI in equation (7) can have a discriminative ability between normal and abnormal conditions. This means that the HI can have different value scales if a machine enters an abnormal stage so that a suitable threshold can be adaptively set to monitor the time of incipient fault initiation based on statistical analysis. For example, if the HI values in a normal stage follow a Gaussian distribution, then a three-sigma rule can be applied to detect incipient machine faults. To eliminate the assumption of Gaussian distribution in a normal stage, an adaptive three-sigma rule (Li et al., 2015) is introduced to determine the time of incipient fault initiation based on a HI. This approach adopts a different trigger mechanism and an incipient fault alarming is activated only when $l+1$ consecutive HI values surpass the three-sigma interval. Herein, l is initialized to 0 and gradually increases from 1 based on iterative steps. The proposed HI construction methodology does not need to use run-to-failure data for model training. The steps for HI construction based on SHAP values for incipient fault detection are summarized as follows.

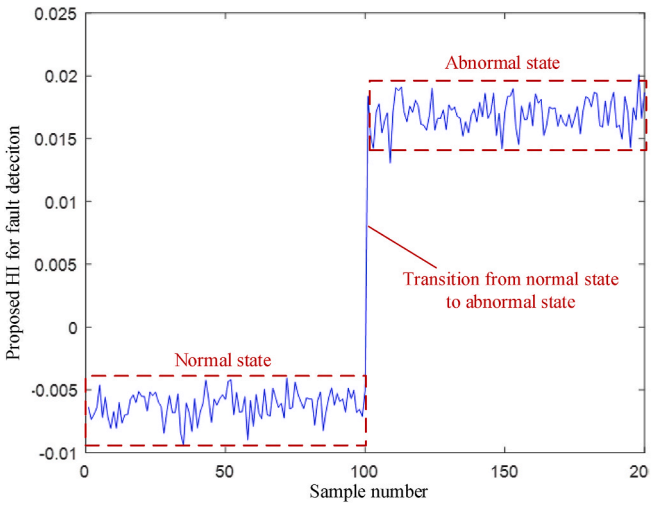

Fig. 7. Generated HI of simulation signals based on SHAP values.

Step 1: Select some normal samples and abnormal vibration samples in a run-to-failure dataset and transform them into spectral amplitudes in the frequency domain;
Step 2: Train a random forest model based on spectral amplitudes of partial normal and abnormal vibration samples for classification;
Step 3: Based on the established random forest model, use the TreeSHAP model to build an explainer;
Step 4: Transform a run-to-failure vibration dataset into spectral amplitudes in the frequency domain;
Step 5: Use the TreeSHAP based explainer to evaluate SHAP values of spectral amplitudes of each data file in a run-to-failure dataset;
Step 6: Based on equation (6), calculate a HI by using SHAP values to monitor machine conditions;
Step 7: Establish a threshold for HI to confirm the time of incipient fault initiation based on the adaptive three-sigma rule (Li et al., 2015).

Based on the above steps, the HI of simulation signals can be generated for anomaly detection as demonstrated in Fig. 7. The proposed HI has a clear mutation trend at the time of fault occurrence for fault alarming. It should be noted that partial normal vibration samples can be selected from the first several data files in a run-to-failure dataset while abnormal vibration samples can be selected from the last several data files in a run-to-failure dataset.

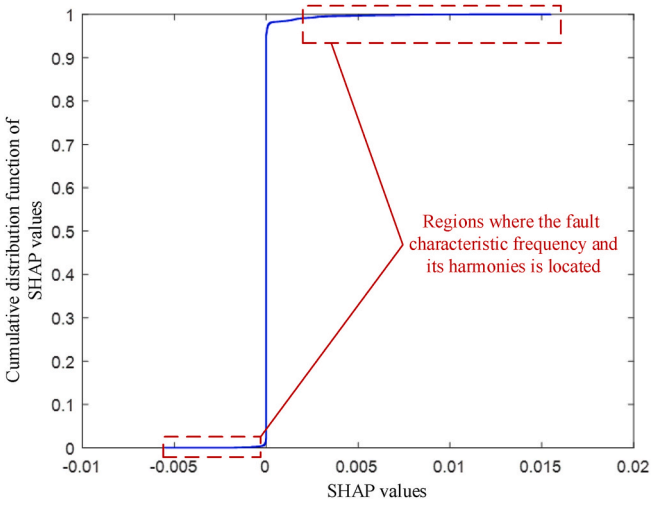

Fig. 8. Empirical cumulative distribution of SHAP values.

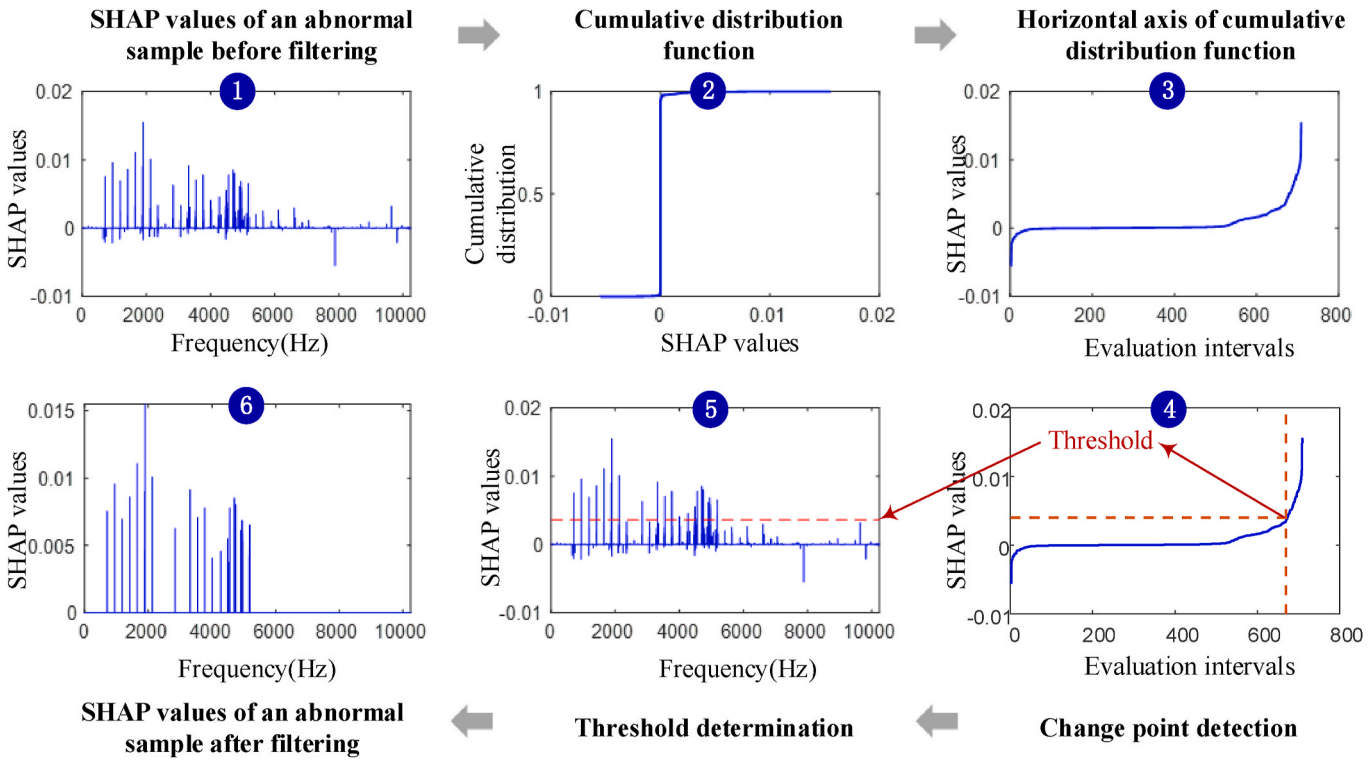

Fig. 9. Procedure for the proposed signal filtering methodology of SHAP values for machine fault diagnosis.

2.2.2. Proposed signal filtering methodology of SHAP values for machine fault diagnosis

Based on the proposed HI, the time of incipient faults can be discovered. The next task is to diagnose faults so that suitable maintenance suggestions can be provided. Although the SHAP values can directly be used as fault characteristics for machine fault diagnosis, they contain many burrs and noise that can obstruct diagnostic performances. Therefore, a signal filtering methodology of SHAP values is introduced in this study so that the cyclic fault frequencies can be revealed by the SHAP values for machine fault diagnosis. In section 2.1, two conclusions of SHAP values of abnormal samples can be drawn. Firstly, peaked SHAP values of spectral amplitudes can be used to locate fault characteristic frequencies in the regions of the natural resonance frequency zone or defect frequency zone. Secondly, most peaked SHAP values of spectral amplitudes are mainly positive among abnormal samples. These two findings are very useful for proposing a signal-filtering methodology for SHAP values and removing irrelevant components, such as noise with low energies in SHAP values.

Inspired by the first finding, the empirical cumulative distribution of SHAP values is first plotted as illustrated in Fig. 8 The empirical cumulative distribution reflects the distribution of data in a ladder pattern. The horizontal axis of Fig. 8 represents evaluation intervals or points of SHAP values that are sorted from smallest to largest while the longitudinal axis of Fig. 8 represents the cumulative distribution function of SHAP values. Since the fault distribution and characteristics of SHAP values are peaked and sparse, they mainly focus on the leftmost or rightmost region of the empirical cumulative distribution as highlighted in Fig. 8. This is because the SHAP values in this region have the highest positive or negative value and their distribution is sparse. However, the second finding indicates the SHAP values of abnormal samples are mainly positive. Therefore, the desirable components of SHAP values in an abnormal condition are mainly located at the rightmost region of the empirical cumulative distribution.

Based on the above analysis, a parametric global model (Lavielle, 2005) is introduced to find a change point in the rightmost region of the empirical cumulative distribution so that an adaptive threshold can be determined automatically to remove other components in SHAP values. This means that if the change point of the empirical cumulative distribution in the rightmost region is found, then the SHAP values that are lower than the SHAP value of the change point are all regarded as irrelevant components and they are removed from the SHAP values. The basic idea of the change point technology is to define a change point as a time instant at which some statistical property of a signal changes abruptly, such as mean or variance.

Given a sequence $\{x_1, x_2, \cdots, x_N\}$ with length N, a parametric global optimization model is established to find a change point k as follows.

$$\min \mathrm{J}(k) = \sum_{i=1}^{k-1} \Delta(x_i; \chi([x_1, x_2, \cdots, x_{k-1}])) + \sum_{i=k}^{N} \Delta(x_i; \chi([x_k, x_2, \cdots, x_N])), \tag{8}$$

where Δ represents the deviation measurement, such as the sum of squared differences. χ means the empirical estimate of the corresponding section.

In this study, the mean is considered to measure its change degree and equation (8) can be transformed as follows.

$$\begin{aligned}\min \mathrm{J}(k) &= \sum_{i=1}^{k-1} (x_i - mean(x_1, x_2, \cdots, x_{k-1}))^2 + \sum_{i=k}^{N} (x_i - mean(x_k, x_2, \cdots, x_N))^2 \\ &= \sum_{i=1}^{k-1} \left(x_i - \frac{1}{k-1} \sum_{r=1}^{k-1} x_r \right)^2 + \sum_{i=k}^{N} \left(x_i - \frac{1}{N-k+1} \sum_{r=k}^{N} x_r \right)^2,\end{aligned} \tag{9}$$

If multiple change points are considered and the number is unknown, adding changepoints invariably leads to a reduction in residual errors. In the worst-case scenario, every single point would be considered a changepoint, ultimately eliminating the residual errors. Therefore, a penalty term (Killick et al., 2012) can be added to the residual errors to avoid overfitting as follows.

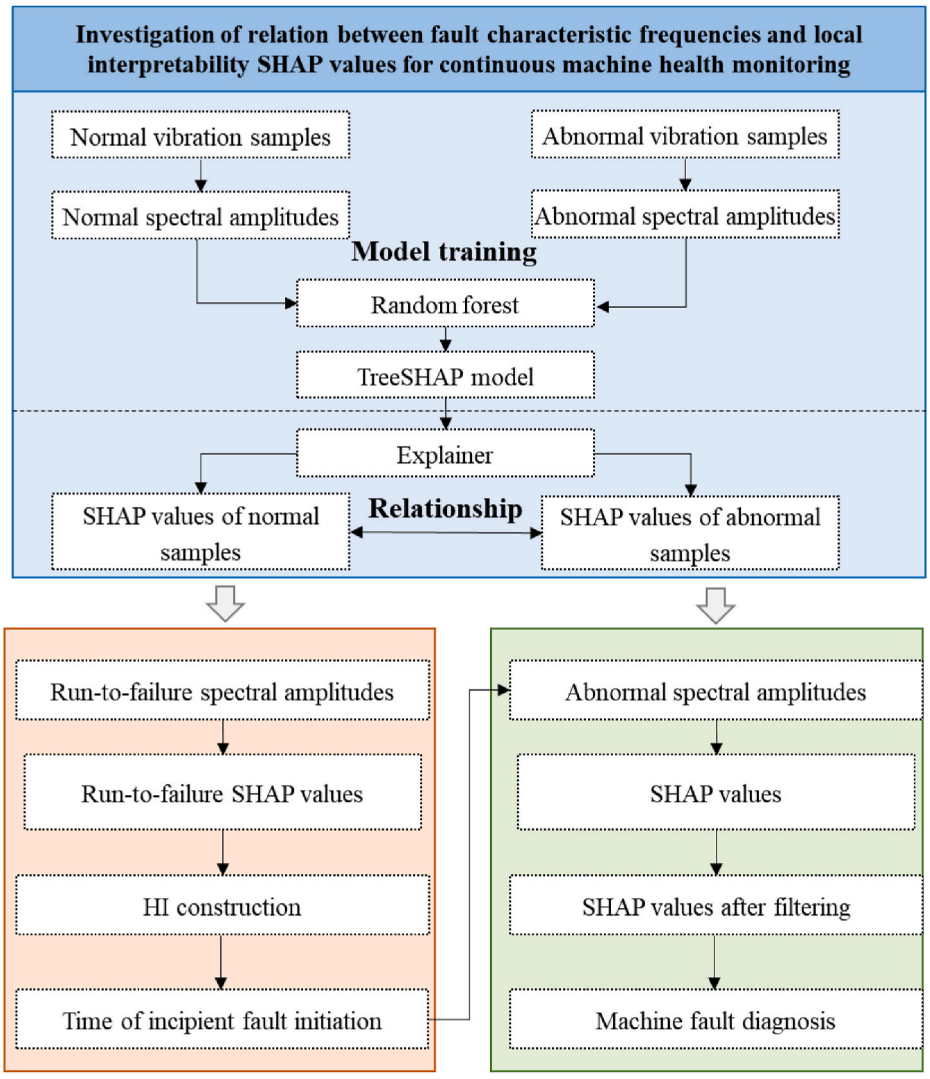

Fig. 10. Investigation on the relation between fault characteristic frequencies and local interpretability SHAP values for continuous machine health monitoring.

$$\min \mathrm{J}(K) = \sum_{r=0}^{K-1} \sum_{i=k_r}^{k_{r+1}-1} \Delta(x_i; \chi([x_{k_r}, \cdots, x_{k_{r+1}-1}])) + \beta K, \tag{10}$$

where k_0 is the first sample of the signal and k_K is the last sample of the signal. β is a proportionality constant. The penalty term in equation (10) associated with these changepoints increases linearly as their number rises.

Based on equation (9), a change point can be detected in the empirical cumulative distribution of SHAP values. Subsequently, the corresponding SHAP value of the change point is used as a threshold to filter out the SHAP values below the threshold. In this way, some desirable components that can show distinguished cyclic fault frequencies can be reserved in the SHAP values while others are automatically removed. The procedures for the proposed signal filtering methodology of SHAP values for machine fault diagnosis are summarized in Fig. 9. Firstly, the empirical cumulative distribution of SHAP values for an abnormal sample can be obtained. A change point detection method is applied to SHAP values of the horizontal empirical cumulative distribution as shown in steps 3 and 4. Subsequently, the SHAP value at a detected change point is regarded as a threshold for signal filtering. Finally, the SHAP values below the threshold are removed to eliminate the noise and unrelated components.

2.3. Whole flowchart of proposed continuous machine health monitoring methodology based on SHAP values

The proposed continuous machine health monitoring methodology based on SHAP values is illustrated in Fig. 10. Unlike most existing studies that only use SHAP values to measure the contributions of different features without further exploration, this study provides a different perspective on using the distributions of SHAP values under different conditions to guide us to propose interpretable methodologies for machine health monitoring. Firstly, the regularity and changes of SHAP values in different health conditions are studied based on a simulation model. It is concluded that peaked SHAP values of spectral amplitudes can reveal fault characteristic frequencies and they have an opposite pattern in normal and abnormal conditions. Based on these useful findings, a HI construction methodology is proposed based on SHAP values for incipient fault detection. Next, a signal filtering method of SHAP values is also proposed to eliminate the burrs and noise in the SHAP values for machine fault diagnosis. This study first gives an insight into the HI construction based on SHAP values. Moreover, some strategies based on change point detection are proposed to process the SHAP values so that they can be used as physics-informed features for physical fault diagnosis. Therefore, this study seamlessly uses the SHAP values to guide us to further put forward some interpretable methodologies for continuous machine health monitoring.

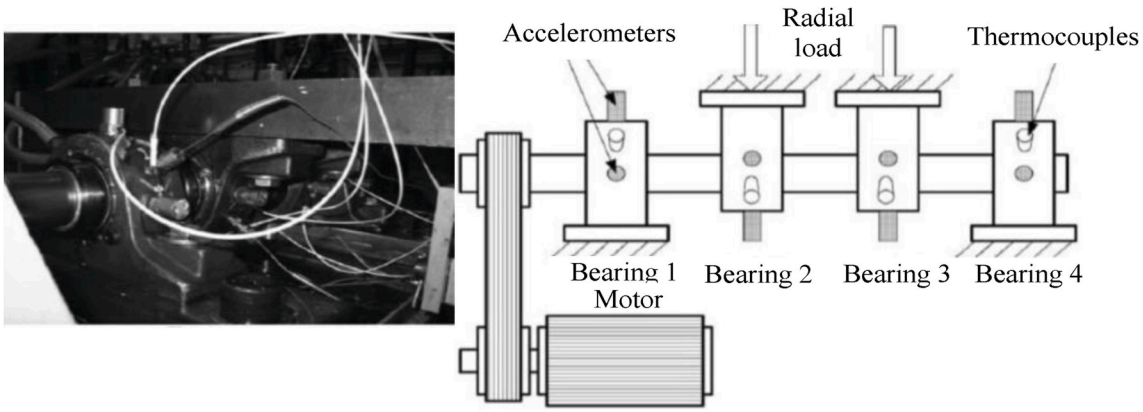

Fig. 11. NASA bearing run-to-failure test platform for vibration data collection (Qiu et al., 2006).

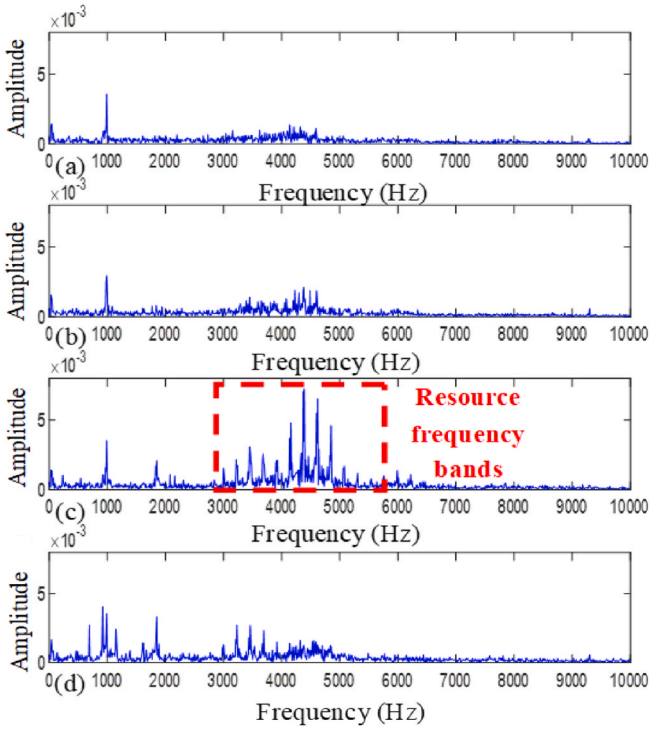

Fig. 12. Spectral amplitudes of different file numbers of bearing 1: (a) file number 100; (b) file number 533; (c) file number 703; (d) file number 900.

3. Results and discussion

In this section, two datasets of vibration signals that were collected in a bearing run-to-failure test platform are used to verify our proposed methodology for continuous machine health monitoring. Furthermore, some existing studies for HI construction and fault diagnosis are introduced and they are compared with the proposed HI construction and fault diagnosis based on SHAP values.

3.1. Proposed HI construction and fault diagnosis methodology based on SHAP values for continuous bearing health monitoring

In the first case study, a vibration dataset collected by the University of Cincinnati is used to demonstrate the effectiveness of the proposed methodology. The endurance test rig of NASA bearing is shown in Fig. 11 and it is composed of four bearings installed in a same shaft and an AC motor. Herein, the AC motor was coupled by a rub belt to maintain a stationary speed of 2000 RPM. During the endurance test of bearing, bearing 1 was confirmed to be invalid and the vibration data was collected under the sampling frequency of 20 kHz. The dataset is composed of numerous files and each file was saved every 10 min, which contains 20,480 vibration samples. At the end of the experiment, a total of 984 files were gathered based on the degradation time of bearing and the timeline of these files records important information about the degradation processes of the bearing from a normal stage to an abnormal stage and finally failed.

The spectral amplitudes of different file numbers are plotted in Fig. 12. The corresponding condition of bearing 1 is in a normal stage in file number 100 and it can be observed from Fig. 12 (a) that only the deterministic frequency component around 1000 Hz is dominant in the frequency domain. It is consistent with Fig. 2 that this frequency component always exists in the frequency spectrum during life cycle degradation processes as shown in Fig. 12(a)–(d). The incipient faults of bearing appear in file number 533 while the frequency responses at this time are very weak. From file number 533 to file number 703, the bearing faults are gradually severe and the fault characteristic frequencies at the natural resonance frequency zone are peaked in this stage. Nevertheless, when the bearing enters a failure stage, the fault characteristics will weaken again as depicted in Fig. 12 (d). Therefore, the frequency amplitudes at different frequency zones evolve with the changes in bearing conditions and they provide abundant information to monitor and diagnose bearing faults. In this study, the relevances between frequency amplitudes and SHAP values are investigated to develop some methodologies for machine health monitoring and diagnosis.

To implement the proposed methodology, vibration samples in each file were first transformed into spectral amplitudes based on FT. All vibration sampling points, namely 20480 of each file, are used to improve the resolution of the spectrum and an effective length of 10240 spectral lines can be obtained because the spectrum is symmetrical. To ensure that the TreeSHAP algorithm can effectively locate fault characteristic frequencies, the signal length at each file used for the FT should be as long as possible to improve the resolution between spectral lines. Therefore, each file is regarded as a sample and 10240 spectral lines of each file are used as useful features in this study. Normal spectral amplitudes from file numbers 1 to 100 and abnormal spectral amplitudes from file numbers 700 to 800 are respectively used to train a random forest for binary classification. After the random forest is established, the TreeSHAP model can be used to generate an explainer. Finally, the SHAP values of 984 files can be acquired and each file has 10240 SHAP values corresponding to 10240 spectral lines at specific frequencies.

The SHAP values of some representative files are given in Fig. 13. In file numbers 100 and 200, the bearing is in a normal stage and the SHAP values in Fig. 13(a)–(b) are mainly located in the fault characteristic frequency and its harmonies in the defect frequency zone and natural

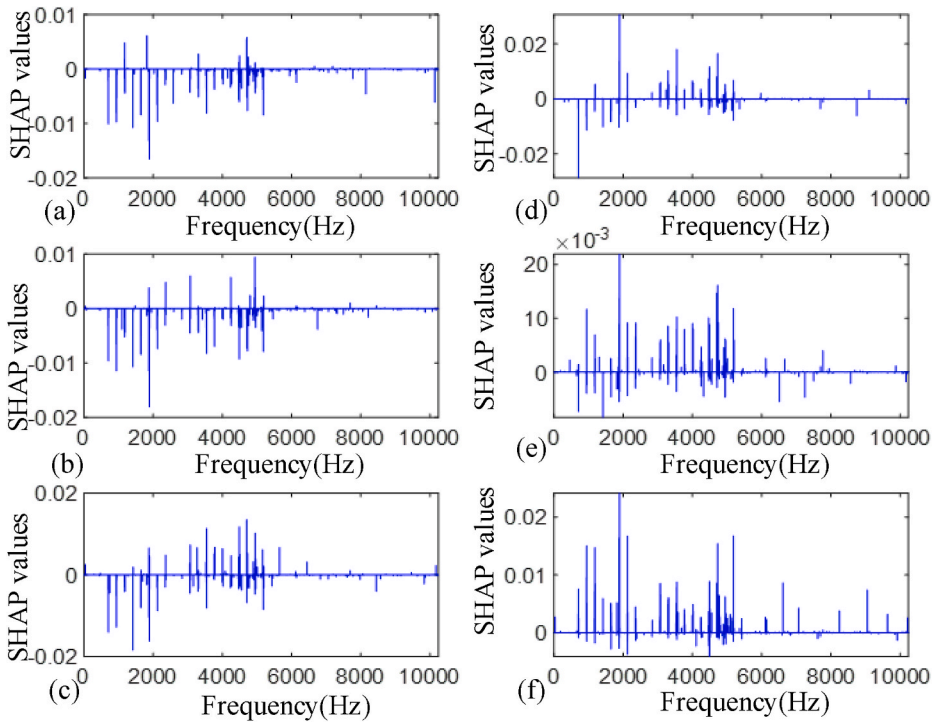

Fig. 13. Obtained SHAP values of different files of bearing 1: (a) file number 100; (b) file number 200 (c) file number 533; (d) file number 550; (e) file number 700; (f) file number 900.

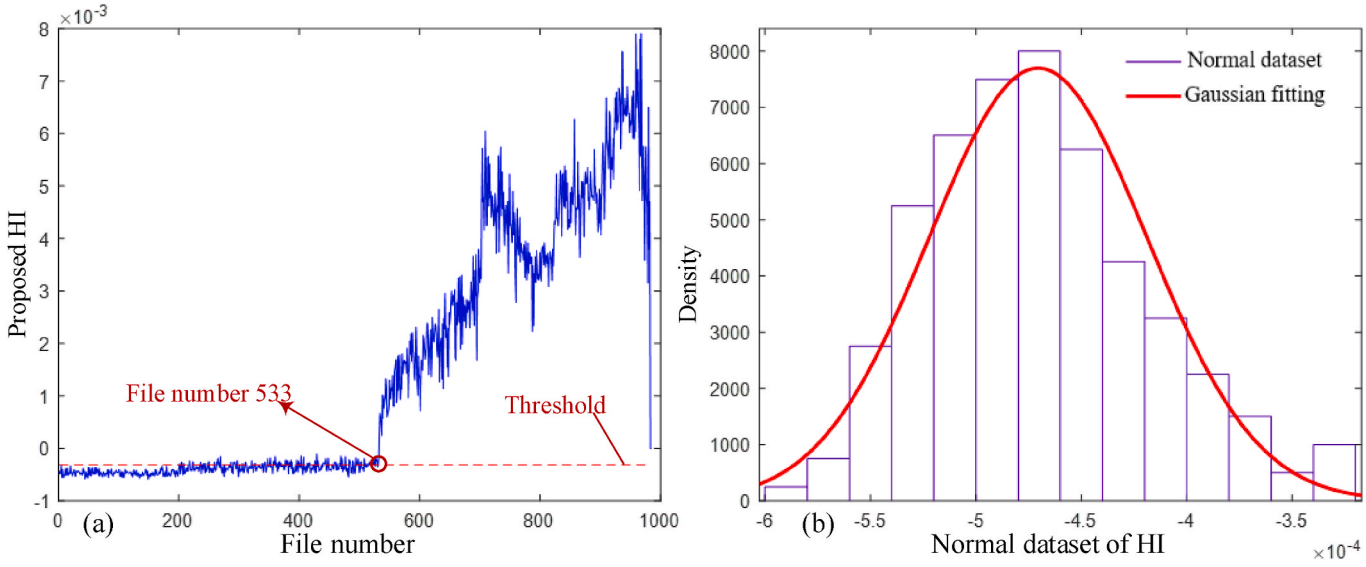

Fig. 14. Proposed HI based on SHAP values for bearing 1: (a) proposed HI; (b) Gaussian fitting of the normal dataset of proposed HI.

resonance frequency zone with negative amplitudes. At the time of incipient fault initiation in file numbers 533 and 534, the bearing is in a transitional stage from a normal condition to an abnormal condition. Because performance degradation of bearing is a slow process and the SHAP values also change gradually from negative amplitudes to positive amplitudes with a slow variation. Therefore, the negative and positive amplitudes of SHAP values approximately account for an equal percentage in file numbers 533 and 534. As the severity of bearing degradation, the SHAP values of file numbers 700 and 900 have large positive amplitudes at fault characteristic frequency and its harmonies. Therefore, it is not accurate to directly use SHAP values for machine health monitoring and diagnosis. It is necessary to propose a more suitable and quantified HI to detect the incipient fault time.

The proposed HI is given in Fig. 14 (a). The first 100 values of the HI in Fig. 14 (a) are used to fit a Gaussian distribution as depicted in Fig. 14 (b). Based on the fitted Gaussian distribution, the mean and standard deviation of the HI under a normal stage can be determined to calculate an initial threshold in Fig. 14 (a) based on a three-sigma rule. The proposed HI has an obvious mutation point at the time of incipient fault initiation of file number 533. To detect the faults quantitatively, an adaptive three-sigma rule is used and the threshold can be adaptively updated for fault time determination. Based on the proposed HI, the time of incipient fault can be discovered in file number 533. The correlation between the established HI and the time waveforms at different bearing

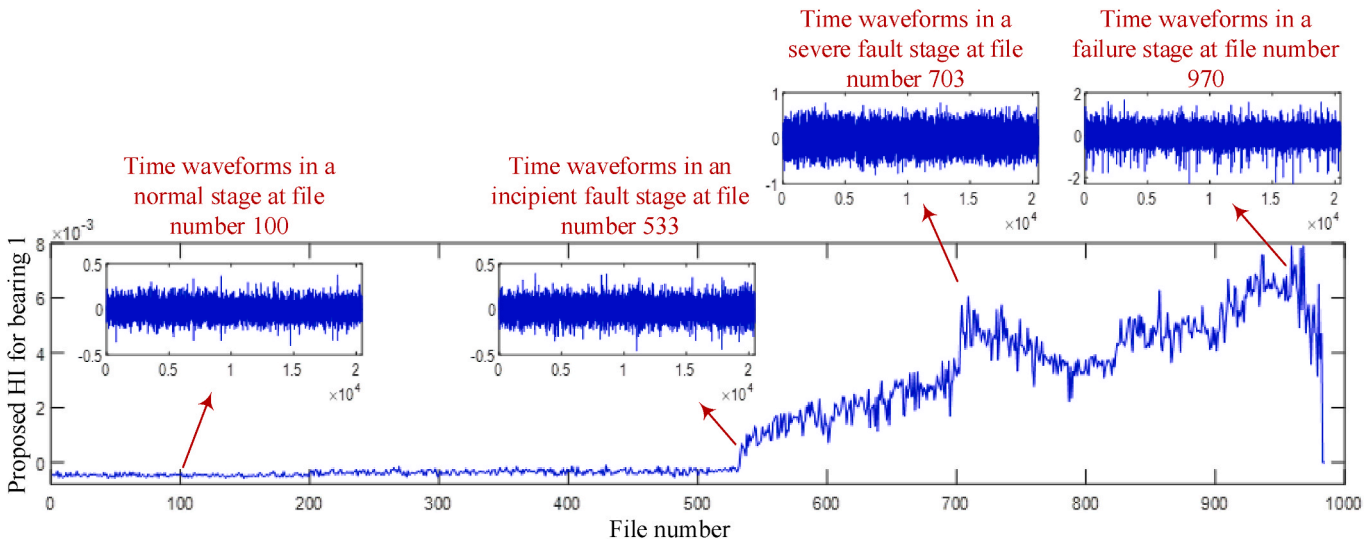

Fig. 15. Correlation between the established HI and the time waveforms at different degradation stages for bearing 1.

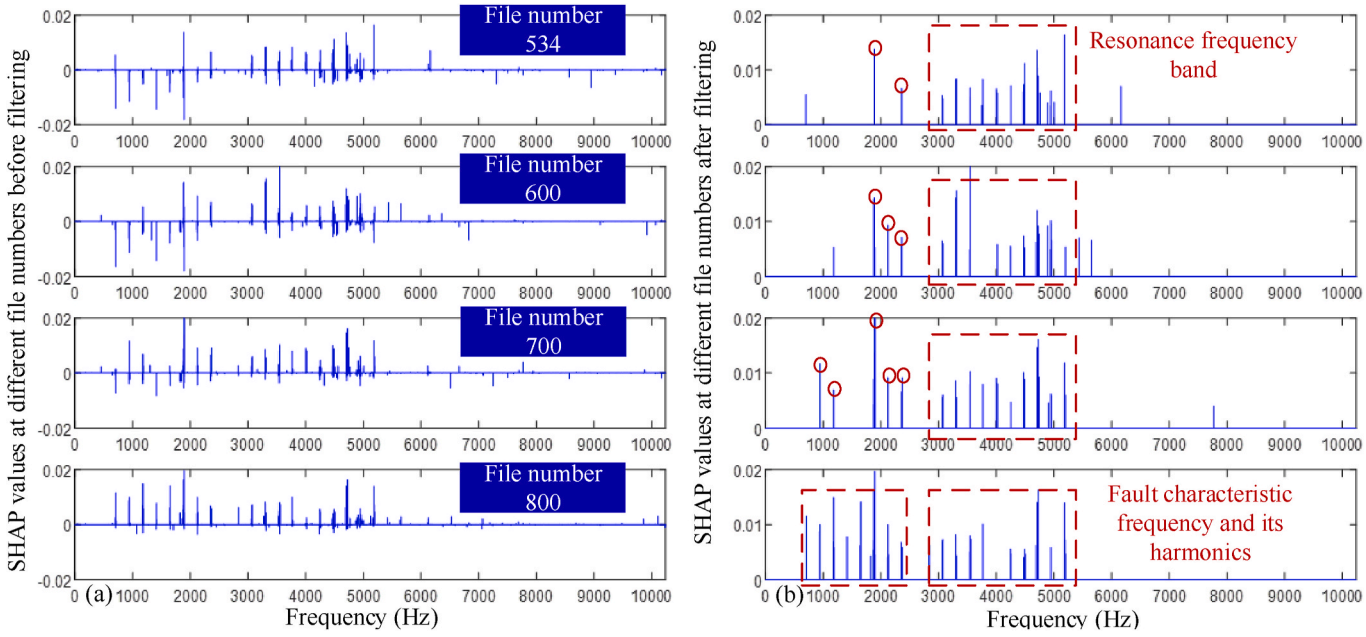

Fig. 16. SHAP values of different file numbers under abnormal conditions before and after filtering: (a) raw SHAP values in file numbers 534, 600, 700, and 800; (b) filtered SHAP values in file numbers 534, 600, 700, and 800 based on the proposed methodology.

degradation stages is demonstrated in Fig. 15, where the fault characteristics of time waveforms in an incipient fault stage at file number 533 are overwhelmed by noise and their amplitudes are similar to the time waveforms in a normal stage at file number 100. Nevertheless, an increasing degradation trend can be indicated by the proposed HI to detect the time of incipient fault initiation.

Subsequently, the fault location needs to be isolated to check the exact fault type. Therefore, the SHAP values of different file numbers under abnormal conditions are filtered based on the proposed methodology so that many burrs and noise can be removed for machine fault diagnosis. The comparisons between the raw and filtered SHAP values in file numbers 534, 600, 700, and 800 are demonstrated in Fig. 16(a) and (b). The SHAP values after filtering based on the proposed methodology can clearly show cyclic fault frequencies in the defect frequency zone and natural resonance frequency zone and their intervals are equal to fault characteristic frequency corresponding to the outer ring. In this way, the SHAP values are physics-informed and they can be directly regarded as significant fault spectrum for machine fault diagnosis. Compared to the original spectral amplitudes in Fig. 12, the SHAP values can show more obvious fault signatures and their amplitudes are more distinguished.

For comparison analysis, some existing HIs are first applied to bearing 1 and their performances for bearing health monitoring are depicted in Fig. 17. Herein, kurtosis (Li et al., 2015) is a typical HI for machine health monitoring and it has been widely used for incipient fault detection. Mobility and complexity are two famous Hjorth's parameters and they are introduced for bearing condition monitoring in a recent work (Cocconcelli et al., 2022). Smoothness index (Chen et al., 2023), negative entropy (Wang et al., 2020a), and Gini index (Liang et al., 2023) are notable sparsity measures and many applications show that they are very sensitive to incipient bearing faults. It can be seen from Fig. 17 that most of them can only show slight changes at the time

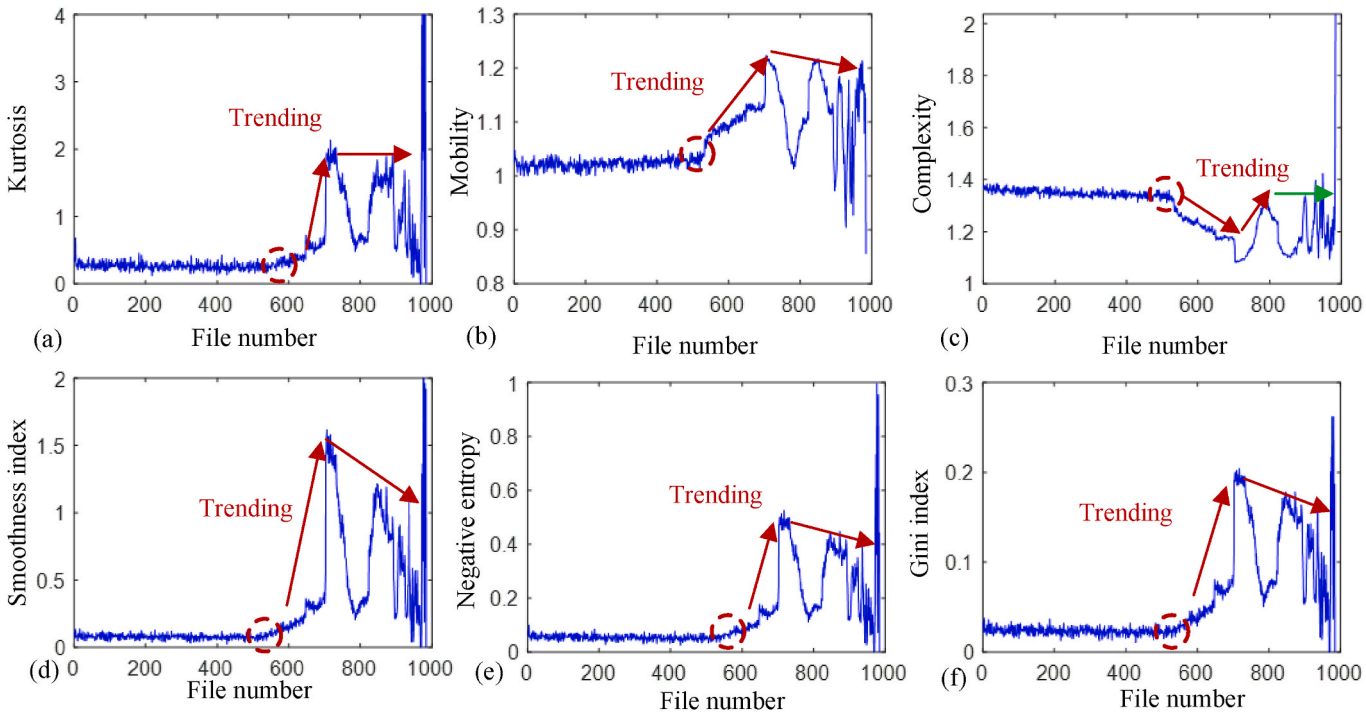

Fig. 17. Exiting HIs for continuous machine health monitoring of bearing 1.

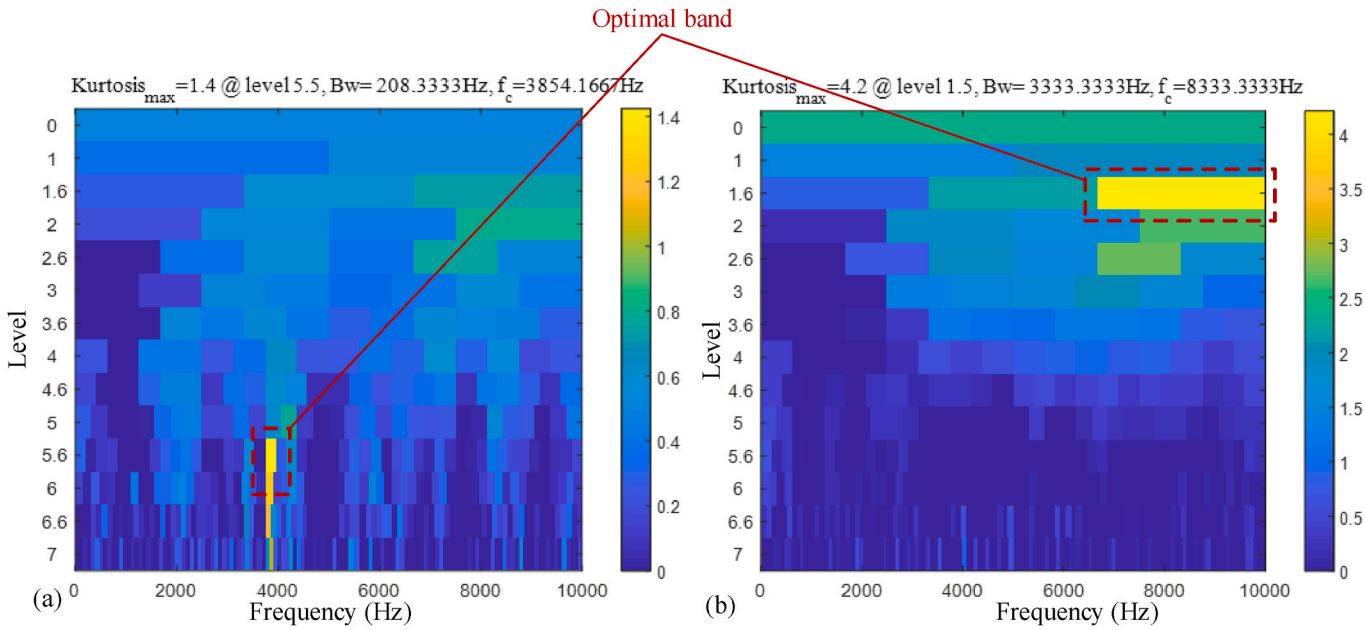

Fig. 18. Informative frequency band detected by the fast Kurtogram at file numbers 533 and 703: (a) file numbers 533; (b) file numbers 703.

of incipient faults compared to the proposed HI in Fig. 14. Mobility and complexity have more clear degradation trends than other HIs while all of them are very fluctuating after the incipient fault time, which are not conducive to fault severity assessment and degradation trend predictions. Although the proposed HI has many burrs and oscillations, it has an overall increasing trend. Therefore, combined with some trend smooth smoothing algorithms, its subsequent monotonic trend after the incipient fault time is promising to be used in fault severity assessment and degradation trend predictions.

Further, a milestone technology for machine fault diagnosis called fast Kurtogram (Antoni, 2007) is also used for comparison. The vibration samples at file numbers 533 and 703 are diagnosed by using the fast Kurtogram and the obtained results are given in Figs. 18 and 19. The indicated informative frequency bands for file numbers 533 and 703 are shown in Fig. 18. The spectral amplitudes of envelope signal filtered by using the informative frequency bands in Fig. 18 for file numbers 533 and 703 are plotted in Fig. 19. It can be observed from Fig. 19 that the fast Kurtogram can diagnose outer ring faults of bearing 1 at file number 703 while it cannot detect the incipient fault time at file number 533. This is because the filtered spectral amplitudes at file number 533 do not show cyclic fault characteristic frequencies. The proposed methodology regards SHAP values as physics-informed features for machine fault

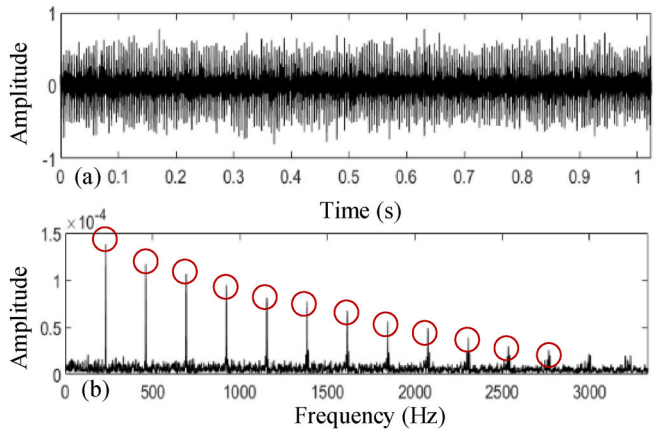

Fig. 19. Filtered spectral amplitudes of the squared envelope by using the informative frequency band detected by the fast kurtogram at file numbers 533 and 703: (a) file numbers 533; (b) file numbers 703.

diagnosis. Combined with the proposed signal filter methodology, the SHAP values can always show clear cyclic fault characteristic frequencies for fault identification. Although the proposed methodology is a data-driven paradigm based on SHAP values, this study associates the SHAP values with some physical elements, such as fault characteristic frequencies for better understanding and interpretability. Moreover, HI construction and fault diagnosis methodologies are accordingly proposed based on SHAP values to realize reliable and interpretable machine health monitoring.

3.2. Proposed HI construction and fault diagnosis methodology based on SHAP values for continuous health monitoring of XJTU-SY bearing

To further demonstrate the performance of the proposed methodology, the test rig for life cycle data collection is plotted in Fig. 20. Several run-to-failure datasets were collected under different working conditions and they ultimately failed due to different types of faults. A dataset labeled as 1-1 is used to verify the proposed methodology and the files in dataset 1-1 were gathered under the rotating speed of 2100 r/min and sampling frequency of 25.6 kHz. The dataset 1-1 contains 123 files and each file has 32767 vibration samples that were saved during the interval of 1.28s. At the end of the endurance test, outer ring failure was found of bearing 1-1 and its theoretical fault characteristic frequency is equal to 107 Hz. Firstly, the spectral amplitudes at different file numbers of bearing 1-1 are shown in Fig. 21. Similarly, the rotating frequency of bearing and its harmonics are always existent during life cycle degradation processes and they are dominant in the normal stage as shown in Fig. 21(a)–(b). The bearing is in an incipient fault stage around file number 78 and a severe fault stage at file number 100. It is difficult to find the fundamental fault characteristic frequency of 107 Hz in the frequency domain, although the amplitudes of its harmonies are gradually increasing. However, the fundamental fault characteristic frequency is the most important frequency line for fault diagnosis.

In this case, all 32767 vibration samples at each file are used to obtain 16384 spectral lines in the frequency domain. Next, these spectral amplitudes of the first 30 files are used as a normal dataset while spectral amplitudes of files 90 to 119 are used as an abnormal dataset for random forest training. A TreeSHAP-based explainer can be established and it is applied to the spectral amplitudes of all files to obtain their SHAP values. The SHAP values of different files under normal and abnormal conditions are demonstrated in Fig. 22. It is interesting to find that the fundamental fault characteristic frequency of 107 Hz can be revealed by SHAP values while it cannot be observed in the original spectral amplitudes. Nevertheless, it contains many burrs and unrelated components during 500 Hz–1000 Hz. Therefore, we can only focus on the frequency components during 0 Hz–500 Hz for machine fault diagnosis. Similarly, it can be observed from Fig. 22(a)–(b) that the SHAP values are concentrated on negative distribution in a normal stage. In a transitional stage in Fig. 22 (c), some positive components in SHAP values will gradually appear and they completely become positive in the fundamental fault characteristic frequency and its harmonies with performance degradation. Therefore, it can further verify the finding proposed in this study about the relationship between the signs of SHAP values and machine conditions.

To accurately determine the time of incipient fault initiation, the proposed HI based on SHAP values is constructed and it is given in Fig. 23. Similarly, the proposed HI can discover the incipient faults of bearing 1-1 at file number 72 based on an adaptive threshold. The correlation between the determined HI and the temporal waveforms corresponding to various stages of bearing degradation is exhibited in Fig. 24. Although the patterns and amplitudes of temporal waveforms at normal and incipient fault stages are similar, the proposed HI still can

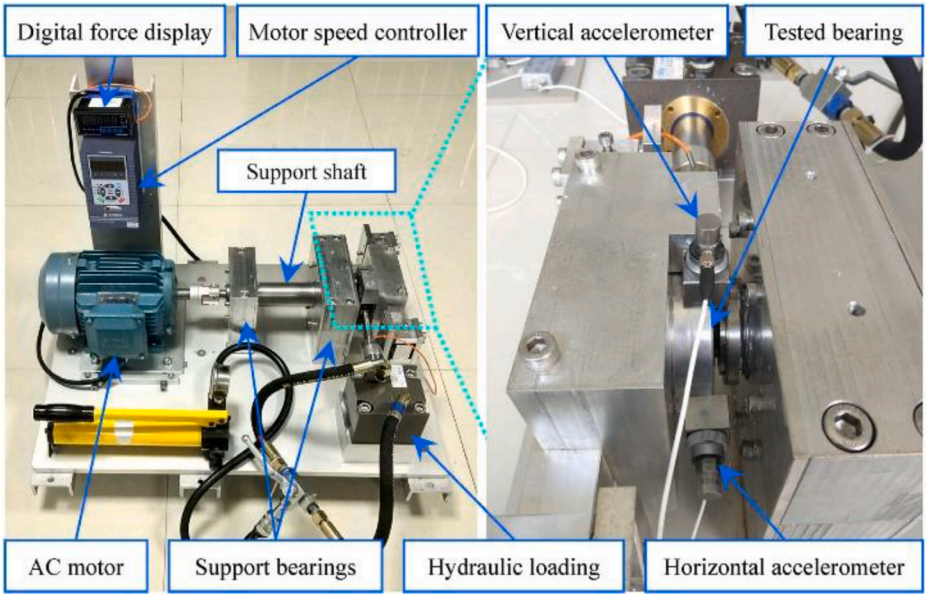

Fig. 20. Test rig of XJTU-SY bearing for run-to-failure vibration data collection (Wang et al., 2020b).

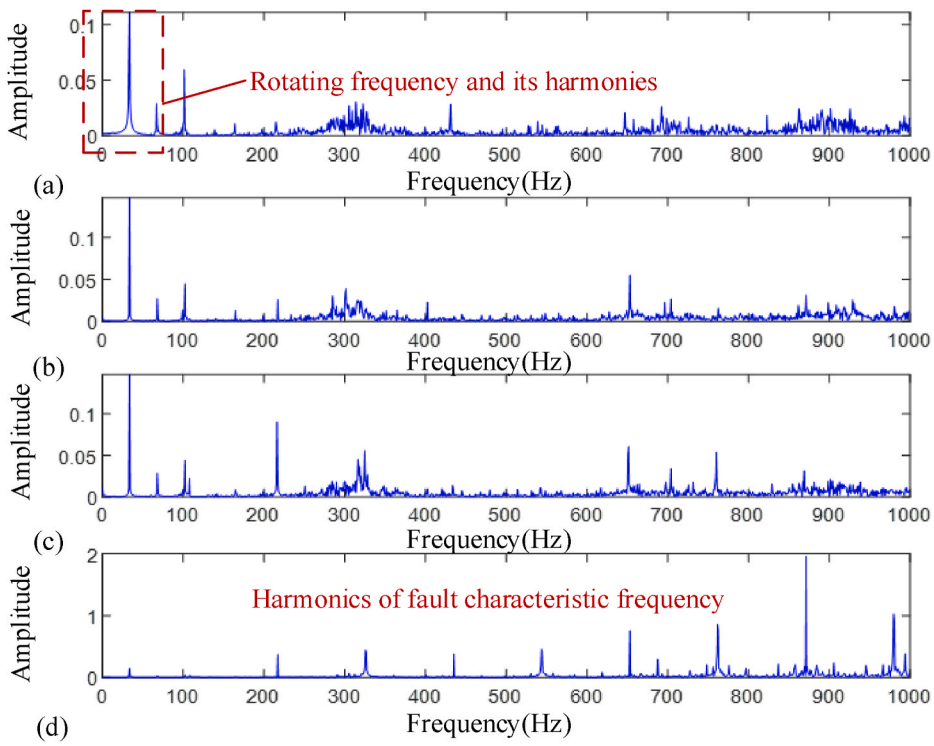

Fig. 21. Spectral amplitudes of different file numbers of bearing 1-1: (a) file number 1; (b) file number 50; (c) file number 78; (d) file number 100.

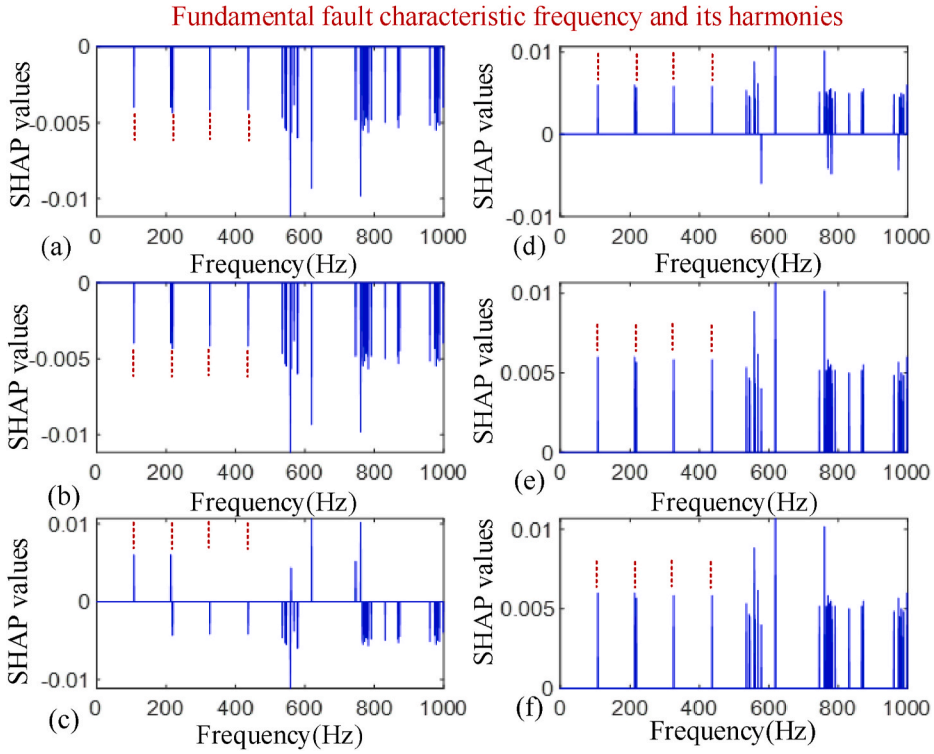

Fig. 22. Obtained SHAP values of different files of bearing 1-1: (a) file number 20; (b) file number 40; (c) file number 70; (d) file number 78; (e) file number 100; (f) file number 120.

show an ability for incipient bearing fault detection.

Subsequently, the proposed signal filtering methodology is applied to the SHAP values under abnormal conditions. The filtered SHAP values at file numbers 80, 100, 110, and 120 between 0 and 50 Hz are demonstrated in Fig. 25. The fault diagnosis of bearing 1-1 can be effectively conducted based on the SHAP values between 0 and 50 Hz, where cyclic fault characteristic frequencies can be clearly revealed. Although there is a noisy component next to the first harmonic of 214 Hz, it does not influence the performances of fault diagnosis based on SHAP values.

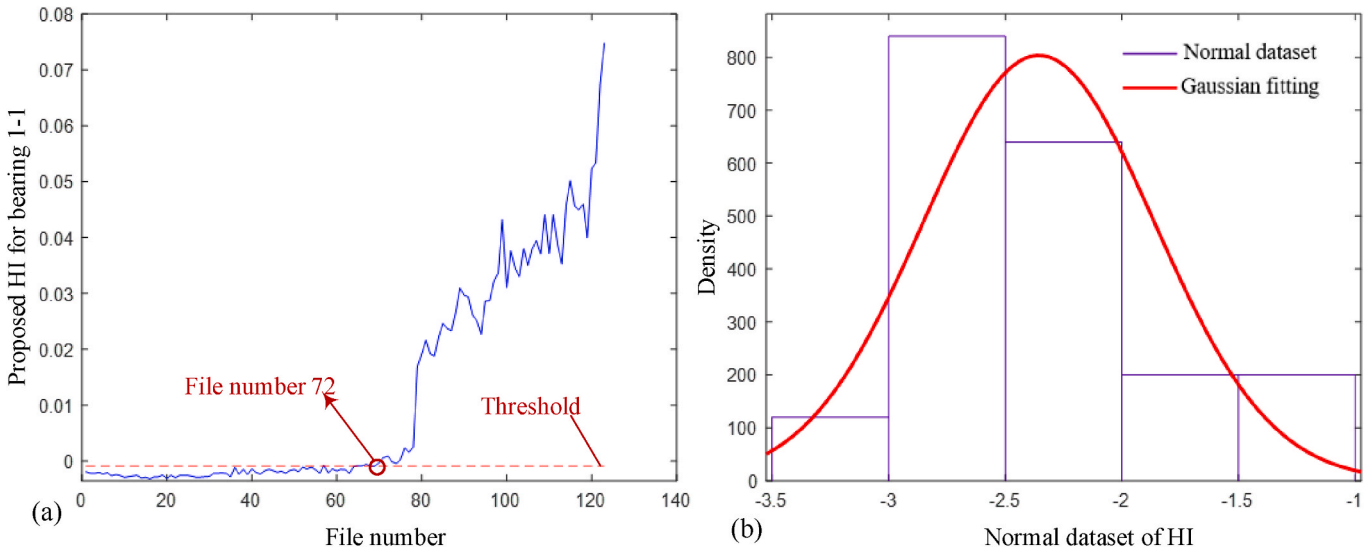

Fig. 23. Proposed HI based on SHAP values for bearing 1-1: (a) proposed HI; (b) Gaussian fitting of the normal dataset of proposed HI.

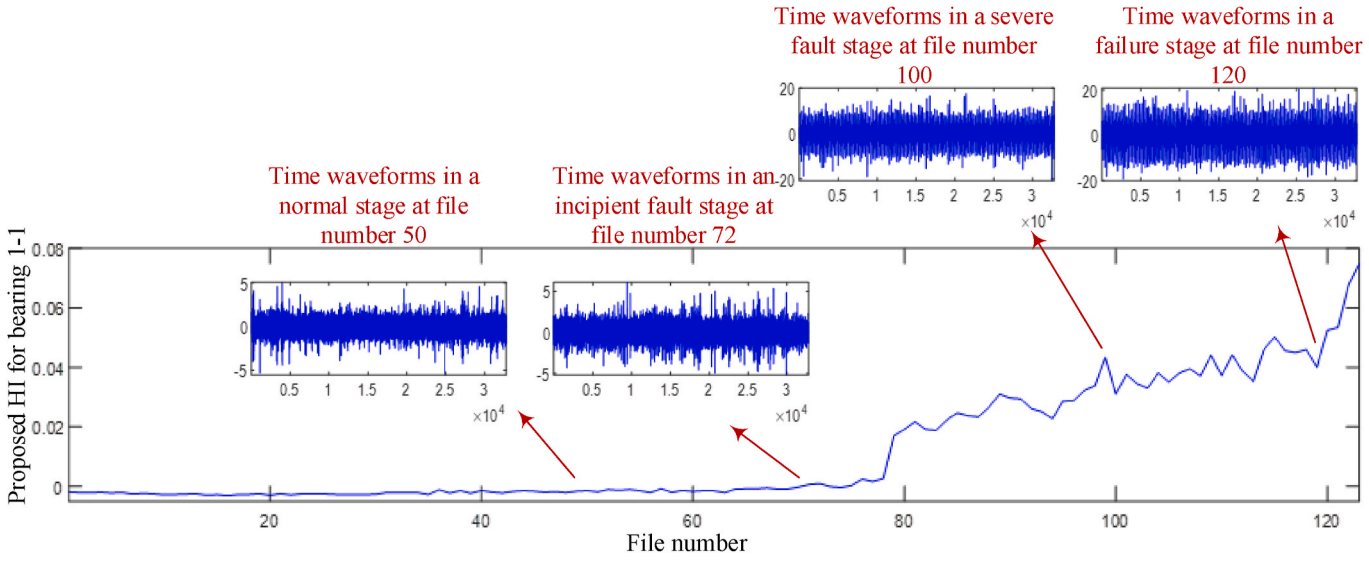

Fig. 24. Correlation between the established HI and temporal waveforms at different degradation stages for bearing 1-1.

Finally, for comparison analysis, existing HIs including kurtosis, negative entropy, Gini index, smoothness index, mobility, and complexity for health monitoring of bearing 1-1 are depicted in Fig. 26. Although they can show large value mutation at the time of incipient fault initiation, they show fluctuating and unstable trends in a normal stage. In this way, it is difficult to identify the incipient fault time and determine a threshold for health monitoring. For a desirable HI for machine health monitoring, it is always expected that they can remain stable and have a stable value level in a normal stage. When incipient faults occur, a HI can show increasing, decreasing, or mutation trends for degradation alarming. Therefore, compared to the existing HIs, the proposed HI based on SHAP values is more suitable for bearing health monitoring. Subsequently, the fast Kurtogram is applied to vibration signals of file numbers 72 and 120 for fault diagnosis. The obtained results for information frequency band location and spectral amplitudes of squared envelope based on the filtered signals for file numbers 72 and 120 are respectively given in Figs. 27 and 28. It can be seen from Fig. 28 that the fast Kurtogram fails to identify the fault type at fault samples of file numbers 72 and 120 and the spectral amplitudes cannot show the fundamental fault characteristic frequency and its harmonies. However, the proposed fault diagnosis method based on SHAP values can reveal significant fault characteristics for accurate fault diagnosis. The verification of bearing 1-1 further indicates the effectiveness and superiority of the proposed methodology.

4. Conclusions

This study aimed to investigate the relationship between fault characteristic frequencies and local interpretability SHAP values. Based on SHAP values, bearing HI construction and fault diagnosis methodologies for monitoring incipient faults and immediate fault diagnosis were proposed. Then, a signal filtering methodology of SHAP values was proposed and filtered SHAP values were regarded as physics-informed fault features for interpretable fault diagnosis. Comparison results showed that the proposed HI is more suitable for incipient fault detection and indicated more significant degradation trends than existing HIs. Moreover, filtered SHAP values based bearing fault diagnosis revealed dominant fault characteristic frequencies and their harmonics for robust and interpretable fault identification.

However, some limitations of this study still exist and will be further

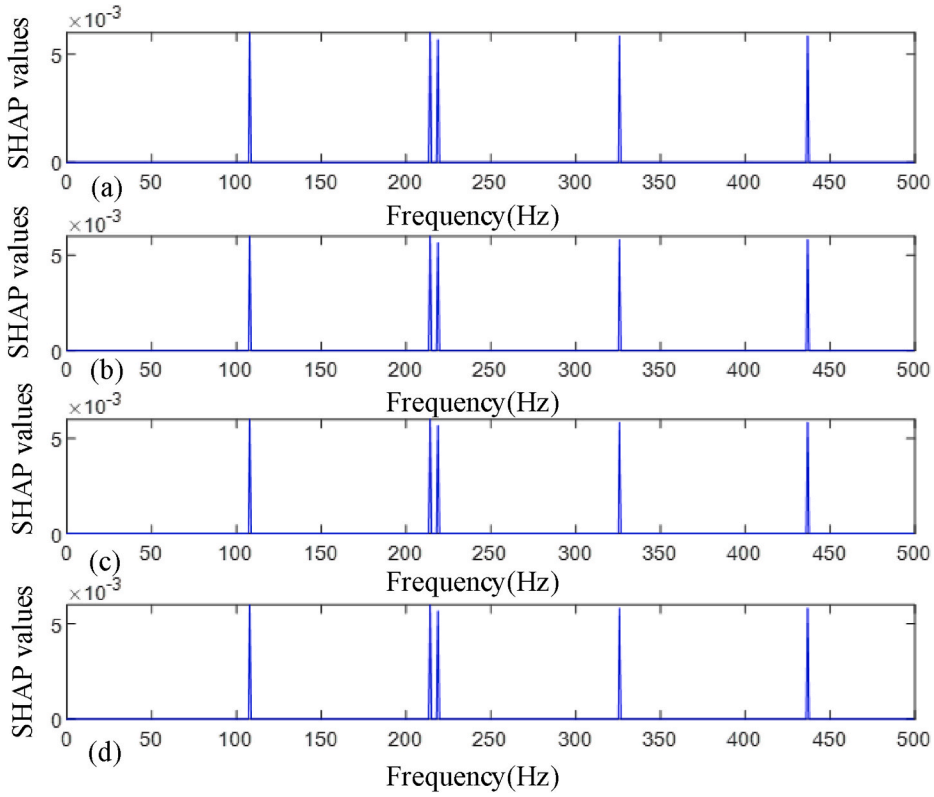

Fig. 25. Filtered SHAP values of different file numbers under abnormal conditions based on the proposed methodology: (a) file number 80; (b) file number 100; (c) file number 110; (d) file number 120.

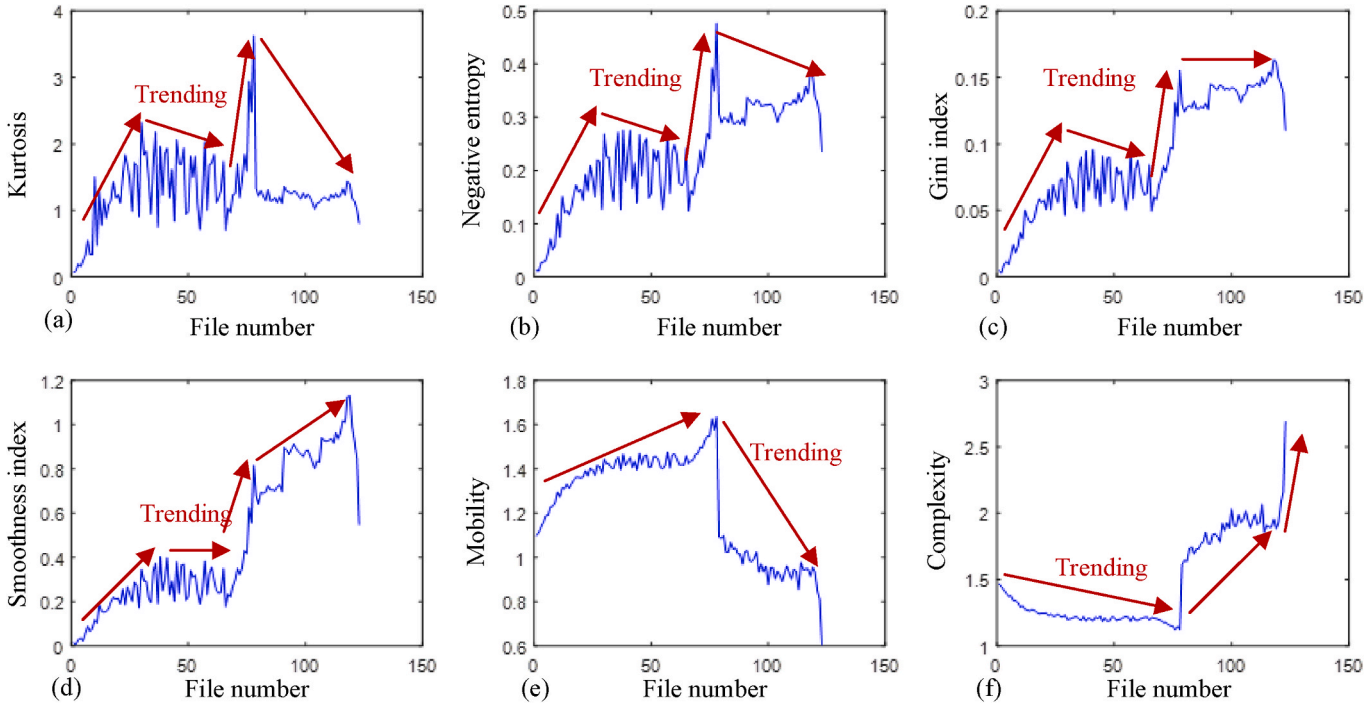

Fig. 26. Existing HIs for continuous machine health monitoring of bearing 1-1.

explored in future. Firstly, the proposed methodology requires fault samples to train a tree model while fault samples are scarce for a new unseen machine or fault. Secondly, this study mainly focuses on local faults in a rotating machine and their fault vibration signals are typically repetitive impulsive signals. The simulation or physical models of signal characteristics for other forms of more complex faults will be further investigated to explore the performance of the proposed methodology. Moreover, the applications of the proposed method for online continuous health monitoring will be explored by using real data from engineering applications. In an online running stage, collected online

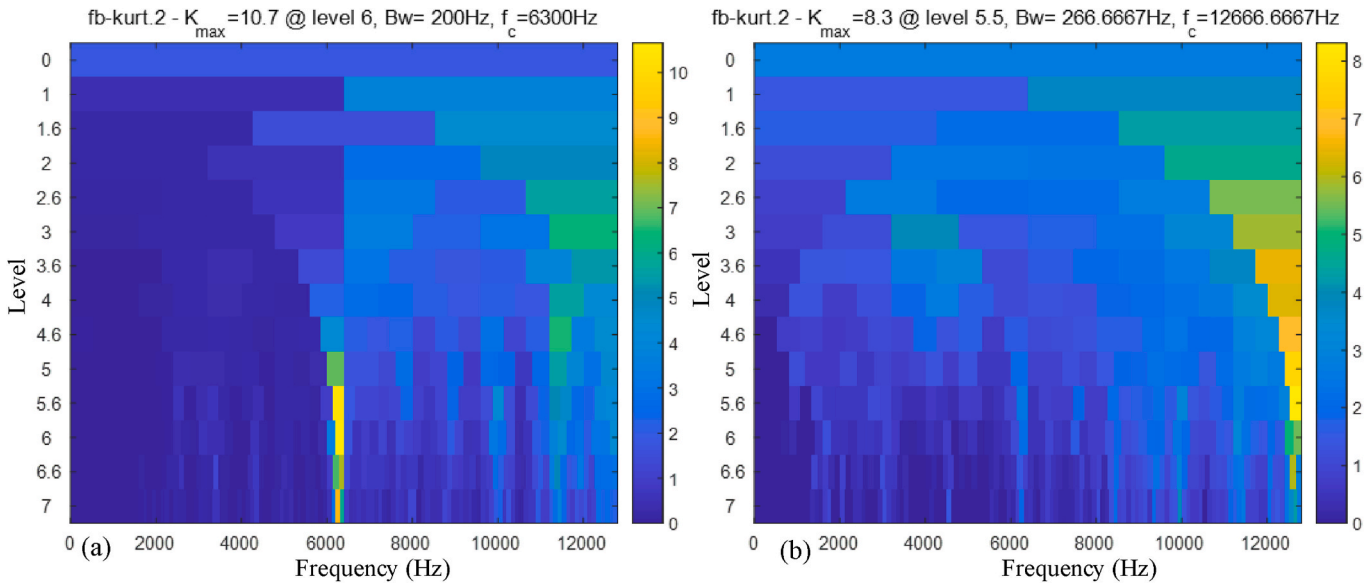

Fig. 27. Informative frequency band detected by the fast Kurtogram at file numbers 72 and 120: (a) file numbers 72; (b) file numbers 120.

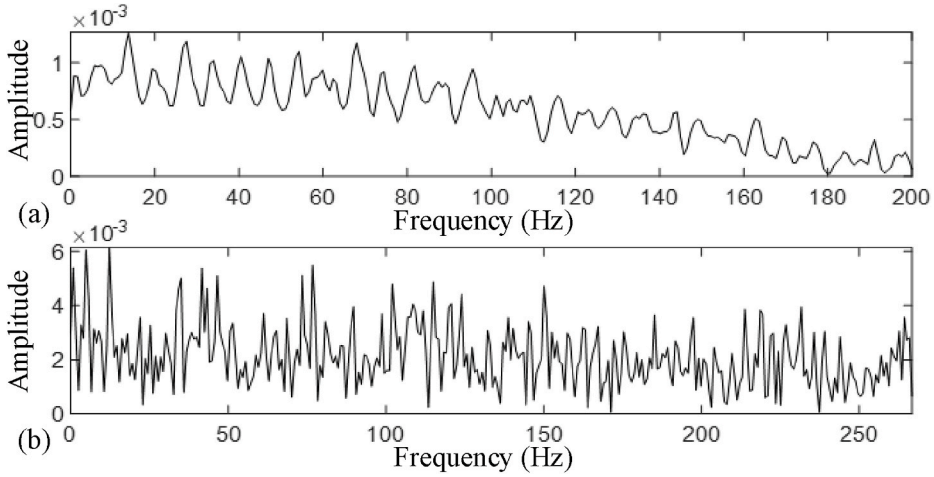

Fig. 28. Filtered spectral amplitudes of the squared envelope by using the informative frequency band detected by the fast kurtogram at file numbers 72 and 120: (a) file numbers 72; (b) file numbers 120.

samples can be first transformed into spectral amplitudes and they are input into the explainer to obtain their SHAP values. Once SHAP values are generated, a HI can be effectively computed for online machine health monitoring. In further studies, the applications of the proposed methodology for other fault modes and agnostic XAI algorithms will be thoroughly investigated.

CRediT authorship contribution statement

Tongtong Yan: Yan, Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Xueqi Xing:** Data curation, Investigation, Validation. **Tangbin Xia:** Supervision, Validation, Formal analysis. **Dong Wang:** Funding acquisition, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The research work was fully supported by the National Key R&D Program of China under Grant No. 2022YFB3402100 and the National Natural Science Foundation of China under Grant No. 12121002.

References

An, B., Wang, S., Qin, F., Zhao, Z., Yan, R., Chen, X., 2023. Adversarial algorithm unrolling network for interpretable mechanical anomaly detection. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2023.3250664.

Antoni, J., 2007. Fast computation of the kurtogram for the detection of transient faults. Mech. Syst. Signal Process. 21 (1), 108–124. https://doi.org/10.1016/j.ymssp.2005.12.002.

Breiman, L., 2001. Random forests, machine learning 45. J. Clin. Microbiol. 2, 199–228.

Chen, B., et al., 2023. Blind deconvolution based on modified smoothness index for railway axle bearing fault diagnosis. In: *Proceedings Of TEPEN 2022*, (Mechanisms and Machine Science, pp. 447–457 ch. (Chapter 38).

Cocconcelli, M., Strozzi, M., Cavalaglio Camargo Molano, J., Rubini, R., 2022. Detectivity: a combination of Hjorth's parameters for condition monitoring of ball bearings. Mech. Syst. Signal Process. 164 https://doi.org/10.1016/j.ymssp.2021.108247.

Decker, T., Lebacher, M., Tresp, V., 2023. Does your model think like an engineer? Explainable AI for bearing fault detection with deep learning. Presented at the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dibaj, A., Hassannejad, R., Ettefagh, M.M., Ehghaghi, M.B., 2020. Incipient fault diagnosis of bearings based on parameter-optimized VMD and envelope spectrum weighted kurtosis index with a new sensitivity assessment threshold. ISA (Instrum. Soc. Am.) Trans. https://doi.org/10.1016/j.isatra.2020.12.041.

Fan, M., Xiao, K., Sun, L., Xu, Y., 2023. Metallogenic prediction based on geological-model driven and data-driven multisource information fusion: a case study of gold deposits in Xiong'ershan area, Henan Province, China. Ore Geol. Rev. 156 https://doi.org/10.1016/j.oregeorev.2023.105390.

Feng, X., Wang, D., Hou, B., Yan, T., 2023. Interpretable federated learning for machine condition monitoring: interpretable average global model as a fault feature library. Eng. Appl. Artif. Intell. 124 https://doi.org/10.1016/j.engappai.2023.106632.

Hoffmann Souza, M.L., da Costa, C.A., de Oliveira Ramos, G., 2023. A machine-learning based data-oriented pipeline for Prognosis and Health Management Systems. Comput. Ind. 148 https://doi.org/10.1016/j.compind.2023.103903.

Jakubowski, J., Stanisz, P., Bobek, S., Nalepa, G.J., 2021. Anomaly detection in asset degradation process using variational autoencoder and explanations. Sensors 22 (1). https://doi.org/10.3390/s22010291.

Jardine, A.K.S., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. 20 (7), 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012.

Jia, F., Lei, Y., Lu, N., Xing, S., 2018. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. Mech. Syst. Signal Process. 110, 349–367. https://doi.org/10.1016/j.ymssp.2018.03.025.

Jin, X., Zhang, X., Cheng, X., Jiang, G., Masisi, L., Huang, W., 2023. A physics-based and data-driven feature extraction model for blades icing detection of wind turbines. IEEE Sensor. J. 1. https://doi.org/10.1109/jsen.2023.3234151, 1.

Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. 107 (500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745.

Lavielle, M., 2005. Using penalized contrasts for the change-point problem. Signal Process. 85 (8), 1501–1510. https://doi.org/10.1016/j.sigpro.2005.01.012.

Li, N., Lei, Y., Lin, J., Ding, S.X., 2015. An improved exponential model for predicting remaining useful life of rolling element bearings. IEEE Trans. Ind. Electron. 62 (12), 7762–7773. https://doi.org/10.1109/tie.2015.2455055.

Li, T., et al., 2021. WaveletKernelNet: an interpretable deep neural network for industrial intelligent diagnosis. IEEE Transactions on Systems, Man, and Cybernetics: Systems 1–11. https://doi.org/10.1109/tsmc.2020.3048950.

Li, Y., Zhou, Z., Sun, C., Chen, X., Yan, R., 2022. Variational attention-based interpretable transformer network for rotary machine fault diagnosis. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3202234.

Li, S., Li, T., Sun, C., Yan, R., Chen, X., 2023. Multilayer Grad-CAM: an effective tool towards explainable deep neural networks for intelligent fault diagnosis. J. Manuf. Syst. 69.

Liang, L., Liu, C., Wang, J., 2023. Periodicity measure of cyclo-stationary impulses based on low sparsity of Gini index and its application to bearing diagnosis. ISA (Instrum. Soc. Am.) Trans. https://doi.org/10.1016/j.isatra.2023.02.017.

Lipovetsky, S., 2022. Explanatory model analysis: explore, explain and examine predictive models. Technometrics 64 (3), 423–424. https://doi.org/10.1080/00401706.2022.2091871.

Lipton, Z.C., 2018. The mythos of model interpretability. Commun. ACM 61 (10), 36–43. https://doi.org/10.1145/3233231.

Liu, W., et al., 2023. Optimal design of γ'-strengthened high-entropy alloys via machine learning multilayer structural model. Materials Science and Engineering: A 871. https://doi.org/10.1016/j.msea.2023.144852.

Lundberg, S., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions.

Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent Individualized Feature Attribution for Tree Ensembles.

Park, J., Kim, S., Choi, J.H., Lee, S.H., 2021. Frequency energy shift method for bearing fault prognosis using microphone sensor. Mech. Syst. Signal Process. 147, 107068.

Park, J.H., Jo, H.S., Lee, S.H., Oh, S.W., Na, M.G., 2022. A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. Nucl. Eng. Technol. 54 (4), 1271–1287. https://doi.org/10.1016/j.net.2021.10.024.

Qiu, H., Lee, J., Lin, J., Yu, G., 2006. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. J. Sound Vib. 289 (4–5), 1066–1090. https://doi.org/10.1016/j.jsv.2005.03.007.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you? Explaining the Predictions of Any Classifier. ACM.

Rocha, E.M., Brochado, A.F., Rato, B., Meneses, J., 2022. Benchmarking and prediction of entities performance on manufacturing processes through MEA, robust XGBoost and SHAP analysis. Presented at the 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA).

Sadoughi, M., Hu, C., 2019. Physics-based convolutional neural network for fault diagnosis of rolling element bearings. IEEE Sensor. J. 19 (11), 4181–4192. https://doi.org/10.1109/jsen.2019.2898634.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. Presented at the 2017. IEEE International Conference on Computer Vision (ICCV).

Shen, S., et al., 2021. A physics-informed deep learning approach for bearing fault detection. Eng. Appl. Artif. Intell. 103 https://doi.org/10.1016/j.engappai.2021.104295.

Wang, D., Peng, Z., Xi, L., 2020a. The sum of weighted normalized square envelope: a unified framework for kurtosis, negative entropy, Gini index and smoothness index for machine health monitoring. Mech. Syst. Signal Process. 140 https://doi.org/10.1016/j.ymssp.2020.106725.

Wang, B., Lei, Y., Li, N., Li, N., 2020b. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. IEEE Trans. Reliab. 69 (1), 401–412. https://doi.org/10.1109/tr.2018.2882682.

Wang, D., Chen, Y., Shen, C., Zhong, J., Peng, Z., Li, C., 2022. Fully interpretable neural network for locating resonance frequency bands for machine condition monitoring. Mech. Syst. Signal Process. 168 https://doi.org/10.1016/j.ymssp.2021.108673.

Xu, W., Fan, S., Wang, C., Wu, J., Yao, Y., Wu, J., 2022. Leakage identification in water pipes using explainable ensemble tree model of vibration signals. Measurement 194. https://doi.org/10.1016/j.measurement.2022.110996.

Yan, T., Wang, D., Zheng, M., Xia, T., Pan, E., Xi, L., 2022. Fisher's discriminant ratio based health indicator for locating informative frequency bands for machine performance degradation assessment. Mech. Syst. Signal Process. 162 https://doi.org/10.1016/j.ymssp.2021.108053.

Yang, Z., Zhang, A., Sudjianto, A., 2021. Enhancing explainability of neural networks through architecture constraints. IEEE Trans Neural Netw Learn Syst 32 (6), 2610–2621. https://doi.org/10.1109/TNNLS.2020.3007259.

Yoo, Y., Jeong, S., 2022. Vibration analysis process based on spectrogram using gradient class activation map with selection process of CNN model and feature layer. Displays: Technology and Applications 73.

Yu, Y, Guo, L, Gao, H, et al., 2024. FedCAE: A new federated learning framework for edge-cloud collaboration based machine fault diagnosis[J]. IEEE Transact. Indust. Electron. 71, 4108–4119.

Zhao, M., Lin, J., Xu, X., Lei, Y., 2013. Tacholess envelope order analysis and its application to fault detection of rolling element bearings with varying speeds. Sensors 13 (8), 10856–10875. https://doi.org/10.3390/s130810856.

Zhou, H., et al., 2022. Construction of health indicators for condition monitoring of rotating machinery: a review of the research. Expert Syst. Appl. 203 https://doi.org/10.1016/j.eswa.2022.117297.