Shang *et al.*
Chinese Journal of Mechanical Engineering (2023) 36:9
https://doi.org/10.1186/s10033-023-00838-0

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Denoising Fault-Aware Wavelet Network: A Signal Processing Informed Neural Network for Fault Diagnosis

Zuogang Shang, Zhibin Zhao and Ruqiang Yan*

Abstract

Deep learning (DL) is progressively popular as a viable alternative to traditional signal processing (SP) based methods for fault diagnosis. However, the lack of explainability makes DL-based fault diagnosis methods difficult to be trusted and understood by industrial users. In addition, the extraction of weak fault features from signals with heavy noise is imperative in industrial applications. To address these limitations, inspired by the Filterbank-Feature-Decision methodology, we propose a new Signal Processing Informed Neural Network (SPINN) framework by embedding SP knowledge into the DL model. As one of the practical implementations for SPINN, a denoising fault-aware wavelet network (DFAWNet) is developed, which consists of fused wavelet convolution (FWConv), dynamic hard thresholding (DHT), index-based soft filtering (ISF), and a classifier. Taking advantage of wavelet transform, FWConv extracts multiscale features while learning wavelet scales and selecting important wavelet bases automatically; DHT dynamically eliminates noise-related components via point-wise hard thresholding; inspired by index-based filtering, ISF optimizes and selects optimal filters for diagnostic feature extraction. It's worth noting that SPINN may be readily applied to different deep learning networks by simply adding filterbank and feature modules in front. Experiments results demonstrate a significant diagnostic performance improvement over other explainable or denoising deep learning networks. The corresponding code is available at https://github.com/albertszg/DFAWnet.

Keywords Signal processing, Deep learning, Explainable, Denoising, Fault diagnosis

1 Introduction

Condition-based maintenance (CBM) of machinery is important in the modern industry [1]. Failure of major equipment, such as aero-engine and helicopter, might lead to huge losses of life and property. While operating in complex conditions, key components of these machines will deteriorate over time [2–4]. Therefore, it's essential to develop the CBM system and detect faults accurately [5]. As the key component of CBM, effective fault diagnosis can reduce the risk of unplanned shutdown [6].

Different signal processing (SP) techniques have been developed and widely used for fault diagnosis tasks [7]. The SP-based methods mainly consist of transform-based methods and index-based filtering methods [8]. As a well-known transform-based method, wavelet transform has been applied to fault diagnosis with great progress over the last 20 years [9]. Chen et al. [10] proposed adaptive redundant lifting multiwavelet for compound fault detection. Assisted with quantitative wavelet function selection, Yan et al. [11] proposed an optimized wavelet packet transform for bearing fault diagnosis. As for the index-based filtering method, its core idea is to construct indicators in a low-dimensional space to concentrate diagnostic information contained in original signal

*Correspondence:
Ruqiang Yan
yanruqiang@xjtu.edu.cn
The State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China

with a high-dimensional space. Based on the characteristic of signals, index-based filtering methods construct a health index utilized for detecting fault information in non-stationary signals. Some health indexes, e.g., Gini index, spectral Gini index, adaptive spectral kurtosis, and smoothness index, were proposed for extracting fault features [12]. These SP-based methods have a firm theoretic basis and have been developed in the past few decades for addressing some challenges, such as denoising and weak feature extraction. However, they usually need expert experience and time-consuming featurization that is dependent on both the problem and the dataset.

In recent years, fault diagnosis based on deep learning (DL) has attracted much attention as it's an end-to-end method without the need for expert experience or featurization [13, 14]. The DL-based methods have been successfully applied to different diagnosis tasks. Considering a nonstationary condition, Liu et al. [15] proposed a multiscale kernel based residual convolutional neural network (CNN) to enhance the feature extraction ability. As fault modes are unpredictable in practical applications, Yu et al. [16] proposed DL-based open set fault diagnosis which could reject unknown-class samples by extreme value theory. These DL-based methods are supposed to be used for a CBM decision in a real industrial environment where a wrong decision may lead to unpredictable losses. In such a risk-sensitive field, explainability is essential for users to effectively understand, trust, and manage such a powerful technique [17]. Furthermore, due to strong background noise in the industrial environment and the complicated transmission path of large mechanical systems, fault features are often weak and easily to be submerged. It's imperative that explainable and noise-restrained DL-based fault diagnosis methods should be developed.

SP-based methods are inherently explainable and have been developed for denoising. DL-based methods have strong data-driven parameter learning ability. It's natural to take advantage of both SP-based methods and DL-based methods. Recently, considerable literature has grown up around the theme of designing a DL network combined with some SP-based methods. Based on morphological analysis, Ye et al. [18] constructed a deep morphological operation layer and the extracted features were weighted based on the kurtosis. Utilizing the wavelet transform, Li et al. [19] proposed WaveletKernelNet which replaces the convolution kernel with the wavelet basis. Yuan et al. [20] constructed an interpretable network with a smart lifting wavelet kernel for fault diagnosis. Considering the structure of the extreme learning machine, wavelet transform, Wang et al. [21] proposed a fully interpretable network for locating resonance frequency bands for machine condition monitoring, in which wavelet transform, square envelope, and Fourier transform were incorporated into the fully interpretable network and sparsity measures were used to quantify repetitive transients. Michau et al. [22] proposed a fully learnable deep wavelet transform network for unsupervised monitoring. Inspired by soft shrinkage in denoising, Zhao et al. [23] designed a residual shrinkage module for denoising. However, it still employs the original convolution as the filter. Although these studies perform very well, they haven't provided a simple yet generic perspective to effectively combine the advantages of SP-based methods and DL-based diagnostic methods.

To provide a more general and comprehensive way of integrating SP-based methods with DL-based methods, we propose a new SP informed neural network (SPINN) framework in this paper. To construct this framework, we employ the Filterbank-Feature-Decision (FFD) to provide a comprehensive perspective to unify the literature mentioned above. FFD is a methodology that has been used explicitly or implicitly in machine condition monitoring [8]. Thus, as illustrated in Figure 1, a SPINN consists of a filterbank stage (purple circle), feature stage (pink circle), and decision stage (gray circle). It's notable that the first two stages of SPINN are designed from SP-based methods but the overall framework is realized by DL-based methods. In the concrete design of SPINN, wavelet transform is selected for the first filterbank stage. In the second feature stage, considering the extraction of features in noisy conditions, the thresholding denoising technique is used first. Then index-based filtering method selects the optimal filter to extract fault features. Finally, features are processed and input into the classifier to give a diagnosis decision. As a data-driven neural network, SPINN could learn from data. Meanwhile, benefiting from SP-based methods, SPINN possesses both explainability and other features from SP methods such as noise resistance.

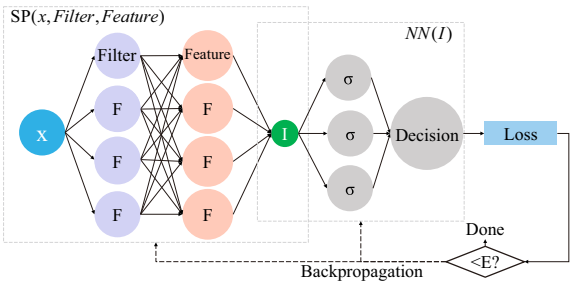

Figure 1 SPINN (Filterbank and feature stages are designed from SP-based methods but realized by DL-based methods. They can extract the information we need (green circle). Then the extracted information is input into a normal neural network (NN) as the final decision stage. The overall SPINN is trained by backpropagation with the target loss until desired training epoch (E))

As a practical DL-based implementation of SPINN, we developed a denoising fault-aware wavelet network (DFAWNet) in this paper. However, like traditional SP-based methods, wavelet denoising and index-based filtering inevitably need expert experience for parameter selection, e.g., selection for wavelet scales, wavelet base types, thresholding function design, filter selection index design, etc. With backpropagation, we try to address these limitations in DFAWNet benefitting from the data-driven mechanism of DL-based methods.

The main contributions of this paper can be summarized as follows.

- Based on the FFD methodology, a framework called SPINN is proposed to effectively take advantage of both SP-based methods and DL-based methods. Although it has been used in the related literatures, SPINN explicitly provides a unified perspective for them. In addition, a concrete SPINN is designed with wavelet denoising and index-based filtering in this paper.
- As a DL-based implementation of the proposed SPINN, the DFAWNet is developed. In the filterbank stage, fused wavelets convolution (FWConv) is designed to implement a learnable wavelet transform with multiple bases. In the feature stage, dynamic hard thresholding (DHT) is proposed to complete hard thresholding corresponding to wavelet denoising. Then index-based soft filtering (ISF) is designed for wavelet filter optimization and selection, so as to extract fault features. In the final stage, the classifier completes the diagnosis decision.
- The proposed SPINN provides an effective way to combine SP-based methods with DL-based methods. In different experiments, DFAWNet shows strong noise resistance ability and fault feature extraction ability by end-to-end learning.

The paper is organized in the following way. Section 2 briefly introduces the main idea of the proposed method and related theory. The DL-based implementation of the proposed SPINN is discussed in detail in Section 3. In Section 4, the method is verified by experiments and comparisons to other methods. Finally, the conclusion is summarized in Section 5.

2 Preliminary

2.1 FFD, SPINN, and DFAWNet

For the SP-based method, a common procedure for fault diagnosis contains filtering, computing health index, and diagnosis. Based on this procedure, the FFD methodology provides a general perspective for the SP-based fault diagnosis method [8]. At the filterbank stage, signals are transformed into different representation domains through a linear filterbank. Then some health indexes are computed for diagnostic feature representation usually with a dimension reduction. Then subband signals with diagnostic features can be selected via these health indexes such as the spectral kurtosis and the energy. Finally, a decision on the health state is taken by manually analyzing selected features or machine learning technologies. Actually, skip of one of these stages is possible.

Based on the FFD methodology, the SPINN is proposed to provide a general perspective for combining SP-based methods with DL-based methods. The core idea of SPINN is shown in Figure 1. For a concrete design of the SPINN, we select the commonly used wavelet transform as the filterbank. To improve robustness to noise and feature extraction ability, thresholding and widely used index-based filtering method (feature frequency band selection method) are used in the feature stage. Finally, these fault features are provided to the classifier and complete the target diagnosis task. The structure is illustrated at the top of Figure 2. For different requirements, SPINN can be designed with different SP-based methods.

As one of the practical DL-based implementations for SPINN, the developed DFAWNet is shown at the bottom of Figure 2. DFAWNet aims to take advantage of the SP-based methods while addressing the problems associated with them. The FWConv aims to alleviate the problem of wavelet basis selection and scale selection in wavelet transform. The DHT solves the threshold function design problem in hard thresholding. The ISF aims to address index design and filter optimization in index-based filtering. Finally, a normal classifier implements the diagnosis task.

In the remainder of this section, we give a brief introduction to the mentioned SP-based methods and their corresponding problems.

2.2 Wavelet Denoising

The procedure of wavelet denoising is illustrated in Figure 3. Since SPINN focuses on feature extraction, we only implement the wavelet transform with thresholding, while not involving the inverse transform in SPINN.

2.2.1 Wavelet Transform (Filterbank)

Supposed that the signal with noise can be formulated as:

$$x(t) = s(t) + n(t), \tag{1}$$

where $x(t)$ is the measured signal; $s(t)$ is the feature signal; $n(t)$ is interference which contains various kinds of noises.

As a widely used signal analysis tool for nonstationary signals, wavelet transform could provide joint information from the time domain and frequency domain.

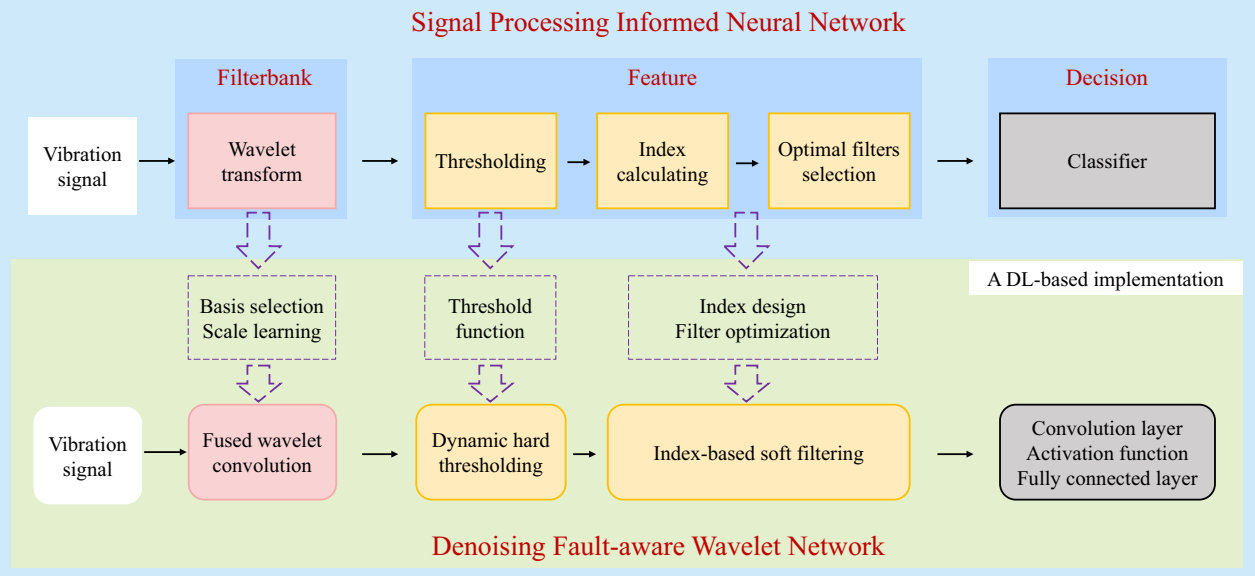

Figure 2 Concrete design of SPINN and its corresponding DL-based implementation: DFAWNet (In the DL-based implementation, we aim to address the problems in the dotted box which are imported by corresponding SP-based methods)

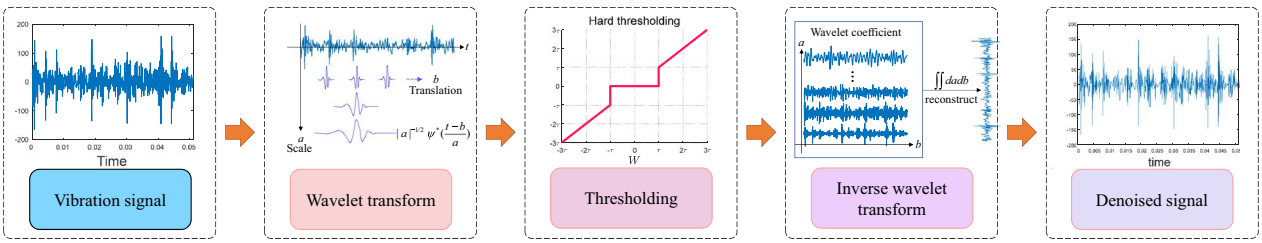

Figure 3 Wavelet denoising

Based on the inner product, the wavelet transform provides decomposition in terms of scale and position, or frequency and time through a series of convolution operations:

$$W_{a,b}(x,\psi) = |a|^{-1/2} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)\mathrm{d}t, \quad (2)$$

where W denotes the wavelet coefficients; a is the scaling parameter inversely proportional to the center frequency; b is the translation parameter to locate the signal; $\psi^*(t)$ is the complex conjugate of the selected wavelet basis $\psi(t)$ from a set of wavelet bases $\Psi = \{\psi^1, \psi^2, \cdots, \psi^N\}$.

The scaling parameter a is the most important parameter which brings the multi-resolution property for the wavelet transform. Different wavelet bases are designed over the past decades for revealing hidden features of non-stationary signals. With the different scales and wavelet bases, wavelet transform can be considered as filters with different characteristics and frequency bands.

Accordingly, choosing a suitable scale and a wavelet basis are fundamental problems when using wavelet transform in the filterbank stage of the SPINN.

2.2.2 *Thresholding (Feature)*

After wavelet transform, the coefficients of $n(t)$ are small and uniformly distributed while those of $s(t)$ are concentrated. Consequently, setting small values to zero is the core idea in thresholding. There are two main types of thresholding methods [24]. Soft thresholding shrinks both negative and positive coefficients towards zero via a threshold, in contrary to hard thresholding which either keeps or removes the values of coefficients. The hard thresholding can be defined as:

$$\hat{W} = \begin{cases} 0, & |W| < \tau, \\ W, & |W| \geq \tau, \end{cases} \quad (3)$$

where $\hat{W}$ denotes denoised wavelet coefficients; τ represents the hard threshold.

Denoising can be seen as a feature selection in each frequency band. To obtain a good denoising performance, the critical problem is to design an appropriate threshold function for wavelet coefficients.

2.3 Index-Based Filtering

Based on prior statistical knowledge of the fault characteristics, index-based filtering selects the optimal filter (frequency band) and extracts fault information. The procedure of common index-based filtering is illustrated in Figure 4.

2.3.1 Filtering (Filterbank)

Firstly, a raw vibration signal is processed by a set of band-pass filters [12]. Here wavelets are used as the filters.

2.3.2 Index Calculating (Feature)

As a key step in index-based filtering, different indexes have been proposed to measure the amount of fault information. In this study, we briefly introduce two types of indicators constructed from energy and sparsity measures.

Energy is a widely used index to find wavelet filters containing defect-related features [25]. The core idea is that the energy of wavelet coefficients is higher in the defect-related frequency band than that of other bands. The energy can be calculated from the corresponding wavelet coefficients:

$$\text{Energy} = \sum_{i=1}^{n} |\hat{W}(i)|^2. \tag{4}$$

In addition, the sparsity measure is widely used to characterize repetitive transients in fault diagnosis. A generalized sparsity measure can be formulated as the sum of weighted normalized square envelop (SWNSE) [26]:

$$\text{SWNSE}_{l,h}\left(\overline{x_{l,h}}[i]\right) = \sum_{i=1}^{N} \text{NSE}_{l,h}[i] \times \varpi[i] - c, \tag{5}$$

where $\overline{x_{l,h}}[i]$ is the modulus (envelope) of the signal $x_{l,h}$ processed by a band-pass filter with a non-dimensional pass-band $l \le k < h$, $\text{NSE}_{l,h}[i]$ is a normalized square envelope (NSE), ϖ is a weight acting on the NSE, c is a constant.

The maximization of the sparsity measure is used to find the optimal band-pass filter. As a typical sparsity measure of the generalized sparsity measure, spectral kurtosis (SK) was proposed in Ref. [27]. Combined with wavelet coefficients after thresholding, it can be expressed as:

$$\text{SK} = \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{W}(i)^4}{\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{W}(i)^2\right)^2} - 2, \tag{6}$$

where $\tilde{W}$ is the envelope of $\hat{W}$:

$$\tilde{W} = \sqrt{\hat{W}^2 + \text{Hilbert}(\hat{W})^2}. \tag{7}$$

SK has the ability to scrutinize the wavelet coefficients on one whole scale, which is related to the central frequency and bandwidth. For the wavelet scale with a higher degree of fault correlation, the value of SK is higher.

Since there are various indexes, designing a suitable index is a fundamental problem when index-based filtering is used in the feature stage of the SPINN.

2.3.3 Optimal Filter Selection (Feature)

The last step is to select the optimal filter based on the designed index:

$$\Theta^* = \arg\max_{\Theta} \text{Index}(x, \Theta), \tag{8}$$

where Θ represents parameters of a set of given filters, e.g., the central frequency and bandwidth; Θ^* are parameters of the optimal filter; $\text{Index}(\cdot)$ is the function to calculate the designed index. For a normal index-based filtering method, Θ is fixed and given in advance, i.e., the frequency band allocation is fixed.

For SPINN, the number of filters is restricted. Thus, the selected filter may not be optimal as the optimal parameters are not contained in the given Θ. However, optimal filter selection is equivalent to the filter parameter

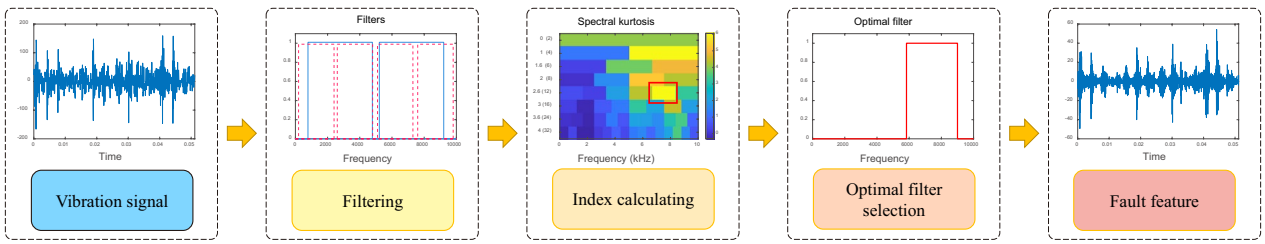

Figure 4 Index-based filtering

optimization problem. The definition of filter optimization is the same as Eq. (8), except for the continuous and optimizable filter parameters. As a result, when index-based filtering is used in SPINN, filter parameter optimization is the second problem we should consider.

3 Proposed Approach

In this section, a detailed introduction to the DFAWNet including FWConv, DHT, ISF, and the end-to-end fault diagnosis network is given below. The core design motivation is illustrated in Figure 2. The overall fault diagnosis framework of DFAWNet is shown in Figure 5.

3.1 Fused Wavelet Convolution

As a DL-based implementation of the wavelet transform, FWConv attempts to address the problem of scale learning and basis selection by the learnable scale and adaptive fusion of various wavelet bases. The learnable scale can locate the appropriate frequency band under the optimization of corresponding loss. Meanwhile, the adaptive fusion of various wavelet bases could effectively improve feature extraction ability. The whole structure of FWConv is shown in Figure 6.

The first step is to realize the wavelet transform with a learnable scale. In Ref. [19], wavelet transform can be inverted into convolution with learnable parameters. Different from Ref. [19], the translation parameter b can be replaced by the stride parameter in 1-D convolution in this study. Consequently, an improved single-parameter wavelet convolution is realized:

$$W_a = x * \psi_a, \tag{9}$$

where a is the learnable parameter of the filter kernel ψ; W_a is wavelet coefficients computed under the scale a. Different channels correspond to different scales. For the sake of convenience, in the cth channel, wavelet coefficients can be denoted as W_c and wavelet basis is denoted as ψ_c.

The second step is to alleviate the problem of basis selection. Different wavelet bases ψ are designed for extracting various features. In addition, a single wavelet basis usually cannot meet the requirement of fault diagnosis [28]. Thus, a simple idea is to prepare a set of wavelet bases and select important ones via the data-driven mechanism.

However, discarding the unselected wavelet bases will stop corresponding gradient backpropagation which is unstable for network learning.

A more proper implementation way for basis selection is to generate C new wavelet bases ψ' from the original ψ based on linear weighting, called fusion:

$$\psi'_i = \sum_{n=1}^{C_o} p_{i,n}\psi_n, \quad i = 1, \ldots, C, \tag{10}$$

where ψ' denotes new wavelet bases; C is the number of generated new wavelet bases; C_o is the number of original wavelet bases ψ; the weight p_i of the ith new wavelet basis satisfies $\sum_{n=1}^{C_o} p_{i,n} = 1$.

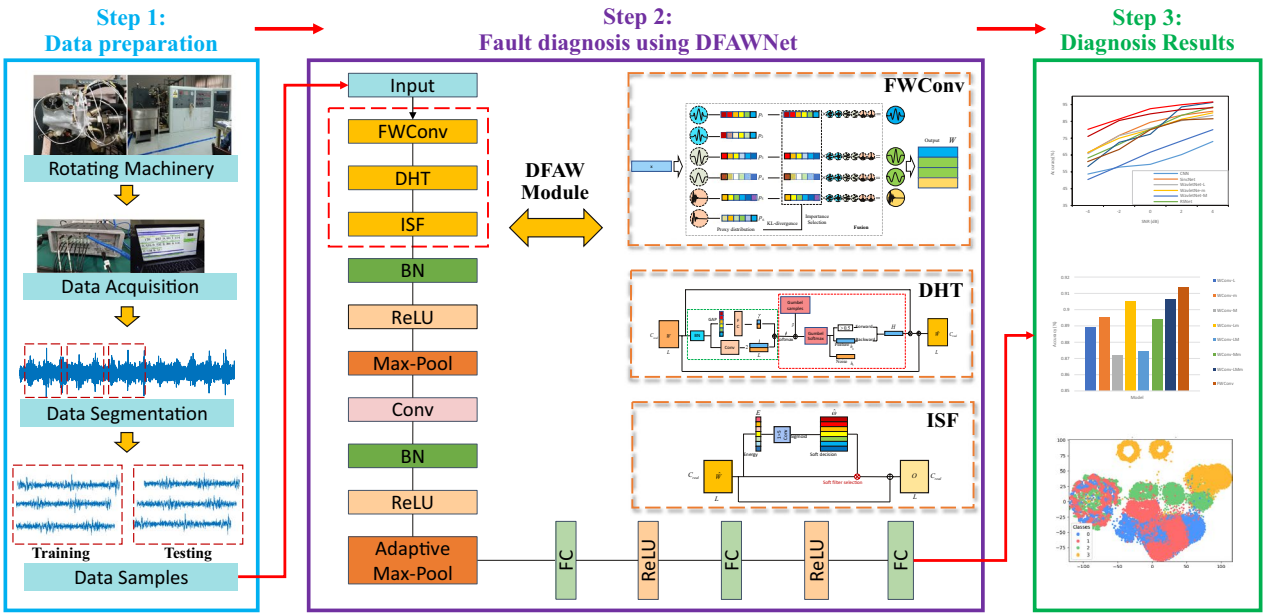

Figure 5 Schematic of DFAWNet for fault diagnosis

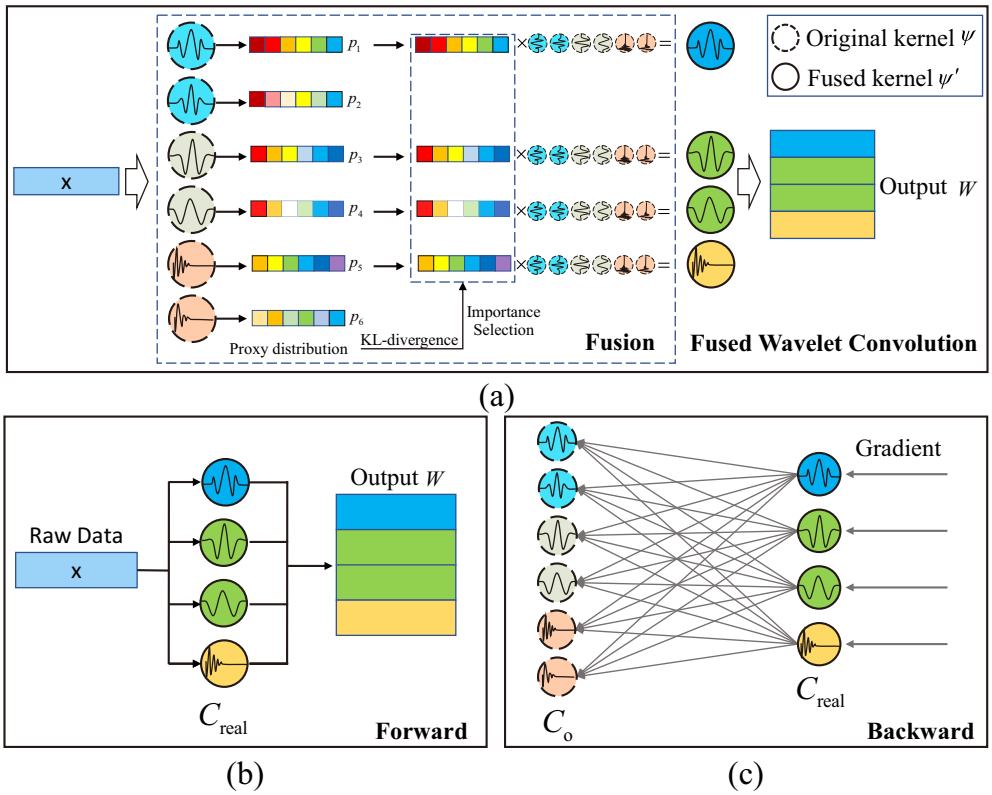

Figure 6 Fused wavelet convolution: **a** New wavelet kernels fused from the original wavelet kernels and used for convolution; **b** Data flow in forward propagation (the fused new kernels are used); **c** Gradient flow in backpropagation (gradient is back propagated to all original kernels)

When the weight p_i is one-hot, we realize the wavelet base selection. For a generalized definition, the weight p_i should be high for the selected wavelet basis ψ_i and low for the basis different from it:

$$p_{i,n} = \text{Softmax}(-t \parallel \psi_i - \psi \parallel_2), \quad n = 1, 2, \ldots C_o, \tag{11}$$

where Softmax$(\cdot)$ guarantees $\sum_{n=1}^{C_o} p_{i,n} = 1$; and t is a temperature parameter changing from 1 to $T = 1e^4$ with the network training.

The basis ψ_i is selected according to the importance index. Inspired by different wavelet bases characterizing various features, a good importance index should be high if the wavelet basis is highly different from each other. Thus, based on KL-divergence [29], the importance index can be defined as:

$$\begin{aligned} I_n &= \frac{1}{C_o} \sum_{m=1}^{C_o} D_{\text{KL}}(p_n || p_m) \\ &= \frac{1}{C_o} \sum_{m=1}^{C_o} \sum_{l=1}^{C_o} p_{n,l} \log \frac{p_{n,l}}{p_{m,l}}, \quad n = 1, 2, \cdots, C_o. \end{aligned} \tag{12}$$

Based on the importance index, we could select C important bases and generate their corresponding weight p. Then we generate new bases and realize the fused wavelet convolution:

$$\begin{aligned} W_i &= \text{FWconv}(x, \psi_i') = x * \psi_i' \\ &= x * \left(\sum_{n=1}^{C_o} p_{i,n} \psi_n \right), \quad i = 1, \cdots, C. \end{aligned} \tag{13}$$

3.2 Dynamic Hard Thresholding

As a DL-based implementation of thresholding, DHT is proposed to address the problem of threshold function design in traditional thresholding methods. Similar to the traditional threshold function, the feature discriminator can give the decision to either keep or remove values of coefficients. Then a reparameterization trick module translates the decision into an optimizable hard thresholding operation. The overall structure of DHT is shown in Figure 7.

According to Eq. (3), the key point of hard thresholding is to determine which coefficients should be kept (feature) and which should be removed (noise). In DL, this

is equivalent to a binary classification problem. Thus, the thresholding can be formulated as:

$$\hat{W} = W \odot H, \quad H = \begin{cases} 0, & \phi < 0.5, \\ 1, & \phi \geq 0.5, \end{cases} \quad \phi \in (0, 1), \tag{14}$$

where $H \in \mathbb{R}^L$ is the operation of removing or keeping and L is the length of the wavelet coefficients; ϕ is the output of the feature discriminator.

The first step is to design the feature discriminator. According to Ref. [30], a threshold function design should consider the inter-scale and intra-scale dependency of the coefficients. For the intra-scale dependency, we should consider the local information of the neighboring coefficient. Then the intra-scale feature is extracted by a simple convolution:

$$[\iota_0, \iota_1] = \text{conv}(W), \tag{15}$$

where $\iota_0 \in \mathbb{R}^L$ is output for the decision to keep the coefficient; $\iota_1 \in \mathbb{R}^L$ is output for the decision to remove the coefficient; conv$(\cdot)$ is a 1-D convolution.

As for the inter-scale dependency, the global average value is used to represent the characteristic of each scale. Then the dependency of different scales is extracted by the fully connected layer:

$$[\gamma_0, \gamma_1] = \text{fc}(\sum_{i=1}^{L} W_{ci} \Big/ L), \tag{16}$$

where $\gamma_0 \in \mathbb{R}^L$ is output for the decision to keep the coefficient; $\gamma_1 \in \mathbb{R}^L$ is output for the decision to remove the coefficient.

Then we obtain the final output of the feature discriminator with the Softmax$(\cdot)$:

$$\phi_i = \text{Softmax}(\iota_i + \gamma_i), \quad i = 0, 1, \tag{17}$$

where $\phi_0 \in \mathbb{R}^L$ represents the decision to keep the coefficients (used as ϕ in Eq. (14)); $\phi_1 \in \mathbb{R}^L$ represents the decision to remove the coefficients.

The next step is to translate the feature decision ϕ_0 into an optimizable hard thresholding operation H. In the inference stage, this feature decision could be directly converted to hard operation by logical judgment. However, in the training phase, directly turning the decision into the hard thresholding operation will result in the loss of gradient for the feature discriminator and probabilistic randomness (probabilistic randomness of the decision is helpful for training when the discriminator is not well trained). These problems can be tackled by the Gumbel-Softmax reparameterization trick [31]:

$$h_i = \text{Softmax}((\log(\phi_i) + g_i)/\varepsilon), \quad i = 0, 1, \tag{18}$$

$$H = \text{re}((1 - \arg\max_i(h_i)) - h_0) + h_0, \tag{19}$$

where g_0 and g_1 are i.i.d samples drawn from Gumbel $(0, 1)$ distribution; ϵ is a parameter that controls the smoothness of distribution h_i, and it's set to 0.66 as this is a binary classification; $(1 - \arg\max_i(h_i))$ returns the sampling result by Gumbel-Softmax; re$(\cdot)$ denotes the reparameterization trick that keeps the gradient backpropagation for h_0. Consequently, in the training stage, H can be denoted as:

$$H = \begin{cases} h_0 > 0.5 \Leftrightarrow 1 - \arg\max_i(h_i), & \text{forward}, \\ h_0, & \text{backward}, \end{cases} \tag{20}$$

where forward denotes the feed-forward stage; backward denotes the backpropagation stage.

Note that dynamic hard thresholding is executed by the optimizable operation H. It provides us with a new perspective to constrain the denoising process from the denoising ratio. The denoising ratio r is defined as the

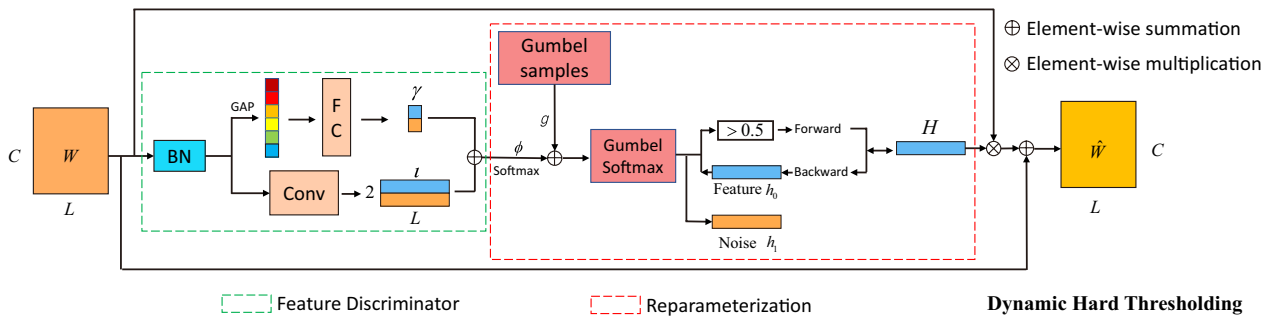

Figure 7 Dynamic hard thresholding module (Signals are dynamically divided into the feature part (should be kept) and the noise part (should be removed) with the feature discriminator. Gumbel-Softmax reparameterization is a trick module that converts feature decisions into an optimizable hard thresholding operation)

proportion of wavelet coefficients that are set to zero in the forward propagation:

$$r = 1 - \frac{\sum H}{L}. \tag{21}$$

A denoising ratio loss L_{DR} is designed for controlling the denoising ratio:

$$L_{\mathrm{DR}} = (r_{\mathrm{actual}} - r_{\mathrm{target}})^2, \tag{22}$$

where r_{actual} is the actual denoising ratio during the training phase; r_{target} is the desired denoising ratio set up previously.

A large denoising ratio implies a strong denoising capability. However, an excessive denoising ratio may eliminate some useful information. Since residual connection can preserve the original signal to stabilize the noise reduction process, the DHT is finally implemented as:

$$\hat{W} = \mathrm{DHT}(W) = W \odot H + W, \tag{23}$$

where the extra identity mapping empirically stabilizes the denoising process and reduces the sensitivity of the hyperparameter setting.

3.3 Index-Based Soft Filtering

As a DL-based implementation of index-based filtering, ISF attempts to address the problems of filter optimization and index designing. Firstly, an index-based loss is constructed for filter optimization. Then the soft filtering selection module selects the optimal filter from those optimized filters based on an adaptive index. The structure of ISF is shown in Figure 8.

Firstly, we design the loss for filter optimization. According to Ref. [27], SK can help in designing more sophisticated filters. Thus, a simple index-based loss can be designed based on Eq. (6):

$$L_{\mathrm{SK}} = -\frac{P(e)L}{C} \sum_{i=1}^{C} \frac{\sum_{l=1}^{L} \tilde{W}_i(l)^4}{(\sum_{l=1}^{L} \tilde{W}_i(l)^2)^2}, \tag{24}$$

where $\tilde{W}$ is the envelope of $\hat{W}$; $P(\cdot)$ is a cos function dynamically changed from 1 to 0 according to the current epoch e.

The index-based loss with the coefficient $P(e)$ enables fast filter optimization via the prior SP knowledge in the early stages of training. Combined with a task-based diagnosis loss, the data-driven mechanism finetunes the filter design in the later stages. After optimization, there are C wavelet filters corresponding to outputs in C channels.

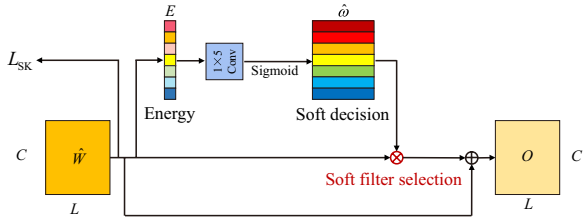

Figure 8 Index-based soft filtering module

The next step is to select the optimal filter from the optimized C filters. Defined as hard filter selection in this paper, traditional index-based filtering often employs only the optimal filter and discards other filters. Each filter has a corresponding output. Thus, filter selection is equal to the corresponding output selection:

$$W_h = \hat{W} \odot \omega, \tag{25}$$

where $\omega \in \mathbb{R}^C$; hard selection ω_i for each channel satisfies $\omega_i \in \{0, 1\}$ and $\sum_{i=1}^{C} \omega_i = 1$.

However, deep learning performs well because it can compose different features of one layer [32]. Instead of hard filter selection, we proposed the soft filter selection to keep all features and implicitly select the optimal filter. In terms of the constraint to ω_i, soft filter selection is a generalization of Eq. (25). Soft selection $\hat{\omega}_i \in (0, 1)$ guarantees that all channels can be combined in the next layer. Furthermore, the condition $\sum_{i=1}^{C} \hat{\omega}_i = 1$ is removed for a more flexible channel selection. In soft selection, a high value of ω_i should corresponds to a channel containing more fault features. Actually, this is what the index should be in the traditional index-based filtering method. Then the problem is to design an appropriate index.

As energy is widely used for constructing a frequency band selection index in wavelet transform, we construct the new index based on the energy. Note that N types of wavelet bases in FWConv lead to N energy characteristics, it's supposed to divide channels into N groups and calculate their relative indices. However, FWConv fuses important wavelet bases dynamically and thus grouping calculation is complicated and resource-consuming. A simple idea is to calculate a local relative index using convolution. Based on the channel energy, the index (soft selection) can be formulated as:

$$\hat{\omega} = \text{sigmoid}(\text{conv}(E)), \tag{26}$$

where sigmoid$(\cdot)$ represents the activation function that scales output value into $(0, 1)$; $E = [E_1, \cdots, E_C] \in \mathbb{R}^C$ represents the channel energy; $\hat{\omega} \in \mathbb{R}^C$ is the index for each channel.

Finally, the ISF is defined as follows:

$$O = \text{ISF}(\hat{W}) = \hat{W} \odot \hat{\omega} + \hat{W}, \tag{27}$$

where $O \in \mathbb{R}^{C \times L}$ are the output features; the extra identity mapping stabilizes the training of the previous part (FWConv and DHT).

3.4 End-to-End Denoising Fault-Aware Wavelet Network Architecture

As shown in Figure 5, DFAWNet is composed of FWConv, DHT, ISF, and a general CNN classifier. We refer to the combination of FWConv, DHT, and ISF as the denoising fault-aware wavelet (DFAW) module. In this paper, the CNN classifier is a relatively shallow 1-D structure modified from Ref. [33].

Firstly, with the noisy raw vibration signal x as the input, the DFAW module extracts more discriminative and robust features based on wavelet denoising and index-based filtering:

$$O = \text{DFAW}_{\nu}(x), \tag{28}$$

where ν denotes learnable parameters of the DFAW module. Then the features are fed into the rest of the DFAWNet, a model $g_\theta(\cdot)$ parameterized by θ, to predict the health state $\hat{y}$:

$$\hat{y} = g_\theta(O), \tag{29}$$

where $g_\theta(\cdot)$ represents the classifier consisting of 1-D convolutional layer, BN layer, ReLU (ReLU is a nonlinear activation function), max pooling layer, and FC layers. The detailed structure is shown in Figure 5.

For a diagnosis task with N_{class} categories, the cross entropy loss is:

$$L_{\text{cls}} = -\sum_{i=1}^{N_{\text{class}}} y_i \log(\hat{y}_i), \tag{30}$$

where y_i is the label of the ith class.

Considering that there are two extra loss functions from DHT and ISF, the total loss for the overall DFAWNet is:

$$L = L_{\text{cls}} + \alpha L_{\text{DR}} + \beta L_{\text{SK}}, \tag{31}$$

where α, β are trade-off parameters, which can be determined by grid search and other hyper-parameter search methods [34].

The DFAWNet parameters are estimated end-to-end by solving the following supervised classification problem:

$$\nu^*, \theta^* = \arg\min_{\nu,\theta} L(\hat{y}, y). \tag{32}$$

As shown in Figure 5, the fault diagnosis framework of DFAWNet consists of 3 steps: (1) Segment acquired signals into fixed-length samples. Divide them into a training set and a test set. (2) Train the DFAWNet with the training set. (3) With the trained network, predict the health state by samples in the test set. As the proposed method only use raw vibration signals as inputs, this is an end-to-end fault diagnosis framework.

4 Experiment Analysis

In this section, experiments on three different datasets are carried out to validate the robustness against noise, generalization ability, and component effectiveness of the proposed DFAWNet respectively. The DFAWNet is implemented by Pytorch 1.10.0 on ubuntu 18.04.6 LTS with NVIDIA GeForce RTX 3090.

4.1 Experiment with XJTU-SY Bearing Dataset

As our target is to realize an explainable denoising model with robustness to noise in the signal, a detailed anti-noise experiment is carried out on this dataset.

4.1.1 Data Description

This dataset is provided by the Institute of Design Science and Basic Component at Xi'an Jiaotong University and the Changxing Sumyoung Technology Co. [35]. The accelerated degradation bearing testbed and corresponding accelerometer layout are shown in Figure 9. Data are acquired at a sampling frequency of 25.6 kHz. The detailed data information is described in Table 1. Similar to Ref. [36], horizontal data at the end of run-to-failure experiments are used. The data from five bearings are divided into five categories. Each signal is segmented into samples of 1024 points without overlap. The ratio of training to testing is 4:1. Thus, there are 512 training samples and 128 test samples.

To validate the robustness to noise of the DFAWNet, Gaussian noise is added to each sample with different signal-to-noise ratios (SNRs) to simulate signals acquired in real industrial equipment. The settings of SNRs are -4 dB to 4 dB with an interval of 2 dB.

4.1.2 Experiment Details

For the sake of comparison, eight models are implemented based on the same backbone, i.e., a baseline model replacing the DFAW module with a normal convolution layer (CNN), an anti-noise model with a wide kernel size of 64 in the first layer (WCNN) [37], a multiscale

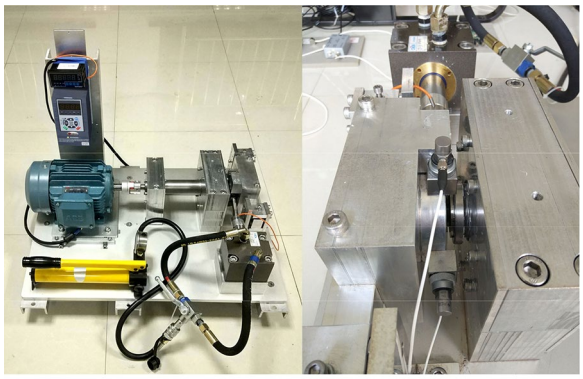

Figure 9 **a** Testbed of rolling element bearings and **b** accelerometer position

kernel model with multi-resolution property (MKCNN) [15], a model with residual shrinkage module (RSNet) [23], an explainable model with a learnable sinc function as the filter (SincNet) [38], three wavelet kernel net with the kernel of Laplace wavelet, Morlet wavelet, and Mexhat wavelet (WaveletNet-L, WaveletNet-m, and WaveletNet-M, respectively) [19].

For DFAWNet, the original channel C_o is set to 128 while the fused channel number C is 64. The wavelet bases are the same as Ref. [19], i.e., Laplace kernel (44 channels), Morlet kernel (42 channels), and Mexhat kernel (42 channels) compose the whole 128 channels. Their scale parameters are initially set as a uniform distribution among different channels, i.e., [0.1, 2] for Laplace, [0.1, 3] for Morlet, and [0.1, 4.5] for Mexhat. Wavelet kernel length is empirically set to 32. The denoising ratio r_{target} is set to 0.2. In the loss function α and β are set to 0.05 and 0.005 respectively. As there's no gumbel noise in the inference phase, it will be no gumbel noise in the last 10% epochs of the training phase to stabilize the inference.

As for the universal training setting, the Adam optimizer with a weight decay of 0.0001 is utilized. The learning rate is 0.0001 with an exponential decay rate of 0.99. The batch size is 64 and the total training epoch is 110. Each experiment is conducted five times to eliminate randomness. These settings are the same in this study unless mentioned otherwise.

Table 1 Detailed description of XJTU-SY datasets

Operation condition	File	Lifetime	Fault element
Speed: 35 Hz Load: 12 kN	Bearing 1	2 h 3 min	Outer race
	Bearing 2	2 h 41 min	Outer race
	Bearing 3	2 h 38 min	Outer race
	Bearing 4	2 h 2 min	Cage
	Bearing 5	52 min	Inner race and outer race

4.1.3 Analysis and Discussion

Figure 10 shows an overview on performance of CNN, SincNet, WaveletNet-L, WaveletNet-m, WaveletNet-M, RSNet, WCNN, MK-CNN, and DFAWNet under different SNRs. In addition, a DFAWNet variant without a DHT module (DFAWNet-1) is conducted to further explore the influence of DHT for enhancing the model's robustness to noise.

It is apparent from Figure 10 that the diagnostic accuracies of all models are reduced with the decrease in SNR. Furthermore, DFAWNet obviously performs better than other models under different noise conditions and shows strong robustness to noise. Besides high accuracies under different SNRs, a relative performance decrease can also show a model's robustness to noise. An index representing noise influence (NI) for the model is then defined as:

$$\mathrm{NI} = \frac{\Delta \mathrm{Acc}}{\Delta \mathrm{SNR} \cdot \mathrm{Acc}_{\mathrm{base}}}, \tag{33}$$

where Δ means difference value between 4 dB and −4 dB; $\mathrm{Acc}_{\mathrm{base}}$ means diagnosis accuracy under 4 dB. A smaller NI represents a lower performance decrease and stronger robustness to noise.

As can be seen from Table 2, DFAWNet obtains the lowest NI which means the best stability of performance among different SNRs. Further analysis shows that the DHT module could effectively improve robustness to noise, e.g., the accuracy improvement from 76.09% to 81.64% under − 4 dB and the IN value decrease from 2.29 to 2.10. What is interesting about the data in this table is that SincNet, WaveletNet-L, and Wavelet-m all have a good IN which means that SP-based methods can help DL to extract more representative features.

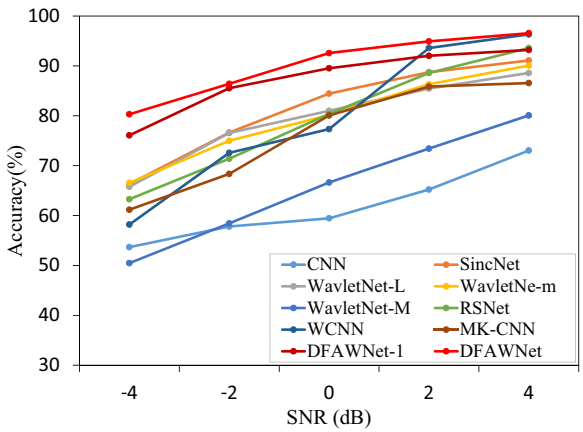

Figure 10 Model performance with different SNRs

As a core part of the denoising process, a more detailed working mechanism of the DHT module is explored. First, we visualize features before the DHT module and after this module in Figure 11. It should be noted that the part of the residual connection is not plotted for a better understanding of the hard thresholding process. We will analyze the residual connection below. As shown in Figure 11, after hard thresholding, it's obvious that some periodic features are kept while other features are set to zero.

The denoising ratio controls how many features are set to set zero. Thus, a high denoising ratio could reduce model performance. To stable the training process and reduce the sensitivity of the denoising ratio, a residual connection is employed in DHT. We conduct a DFAWNet with no residual connection in the DHT module (DFAWNet-NR) and compare it with the original DFAWNet under different denoising ratios. The result of diagnostic accuracy under − 4 dB SNR is illustrated in Figure 12. Without residual connection, a proper setting will obtain better performance. With a denoising ratio of 0.1, the accuracy of DFAWNet is 81.64% while DFAWNet-NR is 81.72%. However, a model with a residual connection could obtain better performance under most settings. In addition, a denoising ratio of 1.0 will lead to a meaningless diagnostic result for DFAWNet-NR and it's not shown in Figure 12.

4.2 Experiment with Aeroengine Bevel Gear Dataset

In the real industrial environment, working condition is not stable and machinery often has rotating speed fluctuation. Thus, DFAWNet is applied to aeroengine bevel gear fault diagnosis under variable operating conditions for illustrating its robustness against operating condition variation.

Table 2 Robustness to noise

Model	Mean accuracy under −4 dB (%)	Mean accuracy under 4 dB (%)	Noise influence (NI)
CNN	53.67 ± 1.36	73.05 ± 2.22	3.32
SincNet	66.33 ± 2.78	91.01 ± 2.52	3.40
WaveletNet-L	65.78 ± 3.98	88.59 ± 2.48	3.22
WaveletNet-m	66.56 ± 2.54	90.08 ± 1.16	3.26
WaveletNet-M	50.47 ± 2.22	80.08 ± 2.38	4.62
RSNet	63.28 ± 2.68	93.59 ± 1.55	4.05
WCNN	58.20 ± 2.4	96.33 ± 0.99	4.95
MKCNN	61.17 ± 2.99	86.56 ± 1.3	3.67
DFAWNet-1	76.09 ± 2.37	93.20 ± 1.78	2.29
DFAWNet	81.64 ± 1.3	96.56 ± 1.0	2.10

4.2.1 Data Description

The lubricating oil accessory testing bench shown in Figure 13 is used for aeroengine bevel gear data collection. With a sampling frequency of 20 kHz, we acquire vibration signals with four different health states, i.e., normal state (NF), tooth surface wear (TSW), broken tooth (BF), and small end collapse (SEC). Considering variable operating conditions, signals under five rotating speeds are collected. Each sample has 1024 points. The detailed information is presented in Table 3.

To compare the model performance between the original operating condition and the cross-operating condition, we set data under 1500 r/min as the original operating condition. The ratio of training data to test data is set to 1:1. There are 4854 training samples with 4850 test samples under 1500 r/min. Data under the other four operating conditions are used as a cross operating condition test set. There is a total of 40149 samples in the cross-operating condition test set.

4.2.2 Experiment Details

Similarly, DFAWNet is compared with CNN, WCNN, MKCNN, WaveletNet-L, WaveletNet-m, WaveletNet-M, DSN, and SincNet in this experiment. Specially, we train and test models at 1500 r/min as a stable operating condition experiment. A robustness test is conducted by training at 1500 r/min and testing on the cross-operating condition test set. The total training epoch is 30.

4.2.3 Analysis and Discussion

The diagnostic results under stable and variable operating conditions are shown in Table 4. From the data, we can see most models can obtain a good accuracy of over 90% and the proposed DFAWNet performs the best.

When these models are applied in variable conditions, there is a dramatic performance decrease for most models. The proposed DFAWNet performs the best in variable conditions and has the smallest accuracy decrease of 14.4% while DSN obtains a performance decrease of 30.63%. In addition, the performance gaps among all models are widened, e.g., the maximum gap among 9 models is 25.03% while it's 10.26% in the stable condition.

In order to illustrate the performance of these models in variable conditions more intuitively, t-SNE is used to visualize the extracted features after the final convolution layer in 2-D spaces (Figure 14). This result is significant that features of SEC (yellow points) are easy to separate for all models. However, CNN, WCNN, MKCNN, WaveletNet, and RSN group features of the other three conditions in the same areas, which leads to poor diagnostic performance. In contrast, as shown in Figure 14(h) and (i), SincNet and DFAWNet separate the remaining

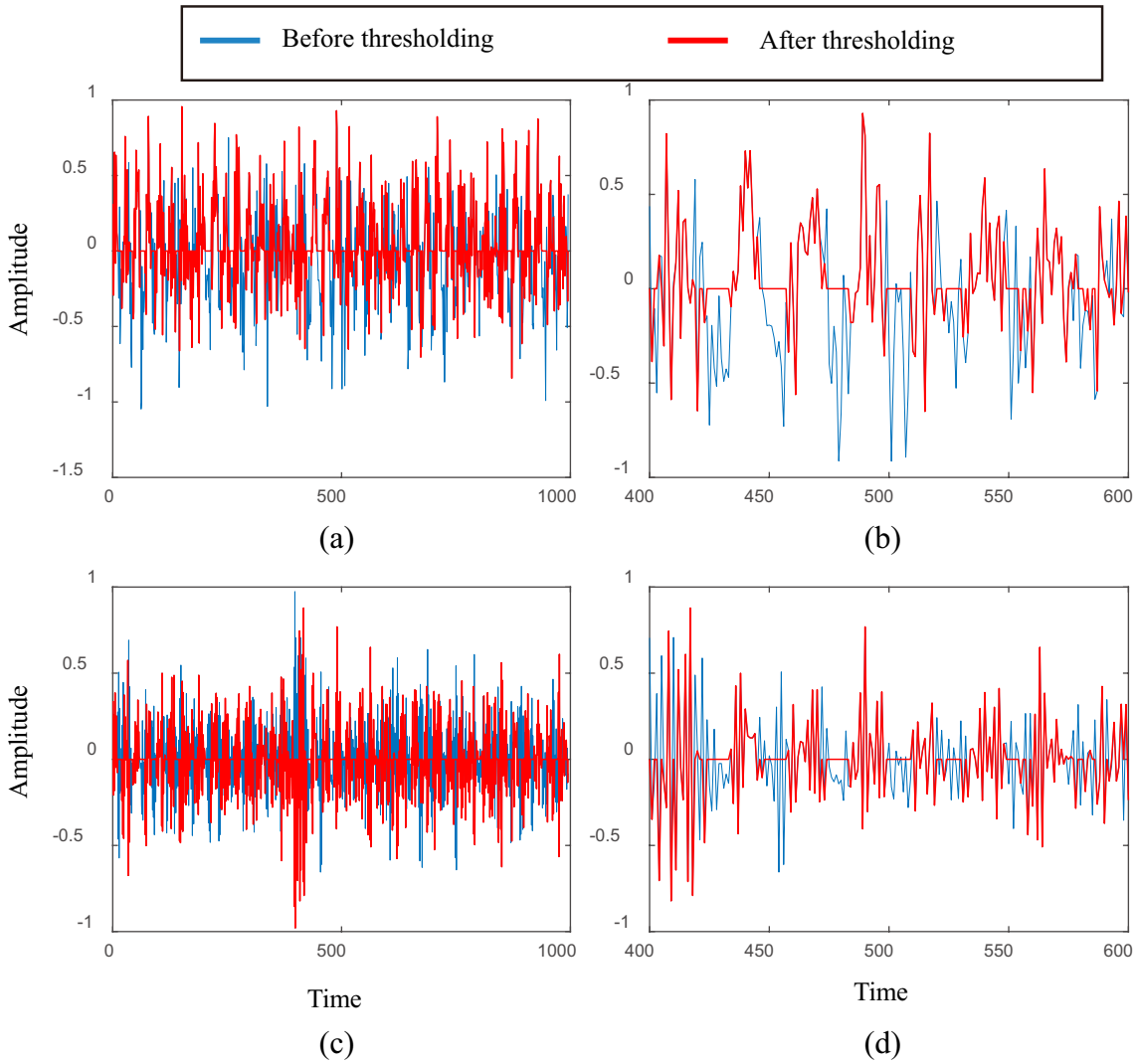

Figure 11 Comparison of features before thresholding and after thresholding: **a**, **b** features of the first channel; **c**, **d** features of the last channel

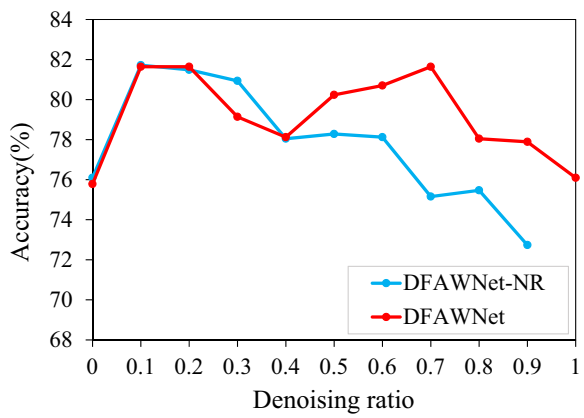

Figure 12 Diagnostic accuracy with different denoising ratios (DFAWNet-NR is DFAWNet without residual connection in the DHT module)

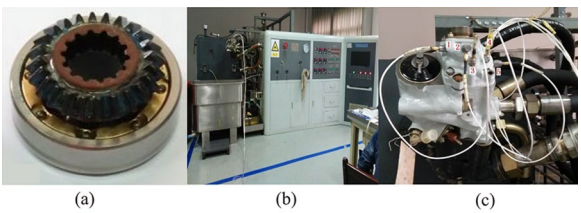

Figure 13 **a** Aeroengine bevel gear; **b** lubricating oil accessory testing bench; **c** sensor position layout

features of the three states into two parts. The features grouped by DFAWNet are more discriminative between state 2 and 3 (green and yellow points). These results suggest that DFAWNet is better than other 8 models in terms of robustness against operating condition variation.

Table 3 Aeroengine bevel gear dataset

Rotating speed (r/min)	Sample number	Health state (label)
500	11537	Normal state (0)
1000	8925	Tooth surface wear (1)
1500	9704	Broken tooth (2)
2000	10805	Small end collapse (3)
3900	8898	

Table 4 Average accuracy (%) with different operating conditions

Model	Stable condition	Variable condition	Decrease
CNN	92.11 ± 2.19	65.69 ± 4.21	26.43
WCNN	94.74 ± 1.42	65.97 ± 6.72	28.77
MKCNN	93.88 ± 0.34	67.33 ± 2.29	26.55
WaveletNet-L	93.34 ± 1.39	65.72 ± 6.08	27.62
WaveletNet-m	87.82 ± 0.47	59.67 ± 1.50	28.14
WaveletNet-M	88.44 ± 2.07	58.65 ± 3.63	29.78
DSN	94.69 ± 1.34	64.06 ± 5.55	30.63
SincNet	97.40 ± 0.23	80.96 ± 3.10	16.45
DFAWNet	98.08 ± 0.05	83.68 ± 1.64	14.40

4.3 Experiment Analysis with MFPT Dataset

As DFAWNet is composed of FWConv, DHT module, ISF module, and CNN classifier, we conduct experiments in this section to verify the effectiveness of each component of the proposed DFAWNet.

4.3.1 Data Description

The Machinery Failure Prevention Technology (MFPT) dataset consists of three sets of bearing vibration data: a baseline set sampled at 97656 Hz, an outer race faults set sampled at 48828 Hz, and an inner race faults set sampled at 48828 Hz [39]. In this paper, we use the baseline data, outer race data at 25 and 50 lbs of load, inner race data at 0 lbs and 50 lbs of load as 5 categories. The sample length is set to 1024. Thus, there are 915 training and 229 test samples.

4.3.2 Experiment Details

The proposed method is first compared with CNN, WCNN, MKCNN, DSN, and SincNet. Then we replace the first layer of CNN with different convolution layers to verify the effectiveness of fused wavelets convolution. WConv-x means a wavelet convolution with wavelet basis x, i.e., L (Laplace), m (Morlet), and M (Mexhat). The channel number is set to 64 and is equally assigned to different wavelet bases if x contains two or three wavelet bases. DFAWNet-1 (DFAWNet without DHT module), DFAWNet-2 (DFAWNet without FWConv), and DFAWNet-3 (DFAWNet without ISF module) comprise the ablation experiment to verify the effectiveness of each component.

4.3.3 Analysis and Discussion

Experiment results shown in Table 5 indicate that DFAWNet performs the best. Comprehensive comparisons of different first convolution layers are shown in Figure 15. These results further confirm that a combination of different wavelet bases can improve model performance, e.g., WConv-LMm is better than other single wavelet basis convolution layers and two wavelet bases convolution layers. It seems that WConv-LM is not better than WConv-L and WConv-Mm is not better than WConv-m. However, we have validated that with a different channel assignment ratio, WConv-LM and WConv-Mm can perform better than corresponding single wavelet basis convolution layers. This is in accord with the result that FWConv is better than WConv-LMm as fusion is actually an adaptive wavelet selection module, which can assign more channels for a proper wavelet basis.

In order to intuitively understand the mechanism of FWConv, we calculate the frequency bands of all wavelet convolution kernels and present the cumulative frequency band of all kernels in Figure 16. The central frequency moves from 6.10 to 12.21 kHz. Furthermore, as shown in Figure 17, the central frequencies of optimal filters locate between 9 kHz and 14 kHz where 12.21 kHz is in the middle. Thus, the learning of FWConv can be regarded as an optimal frequency band adjustment process. It can be explained as concentrating on a frequency containing more fault information.

To verify the effectiveness of FWConv, DHT, and ISF in the proposed DFAWNet, an ablation experiment is conducted and the results are presented in Table 6. Three components are absolutely useful. Further analysis shows that FWConv is the most important part as it's the foundation of wavelet denoising theory which could extract multiscale features. A possible explanation for a relatively small improvement of the DHT module is that the data is acquired in the laboratory with a low-noise environment. Overall, these results indicate that FWConv, DHT, and ISF are important components of the proposed DFAWNet and they can effectively improve diagnostic performance.

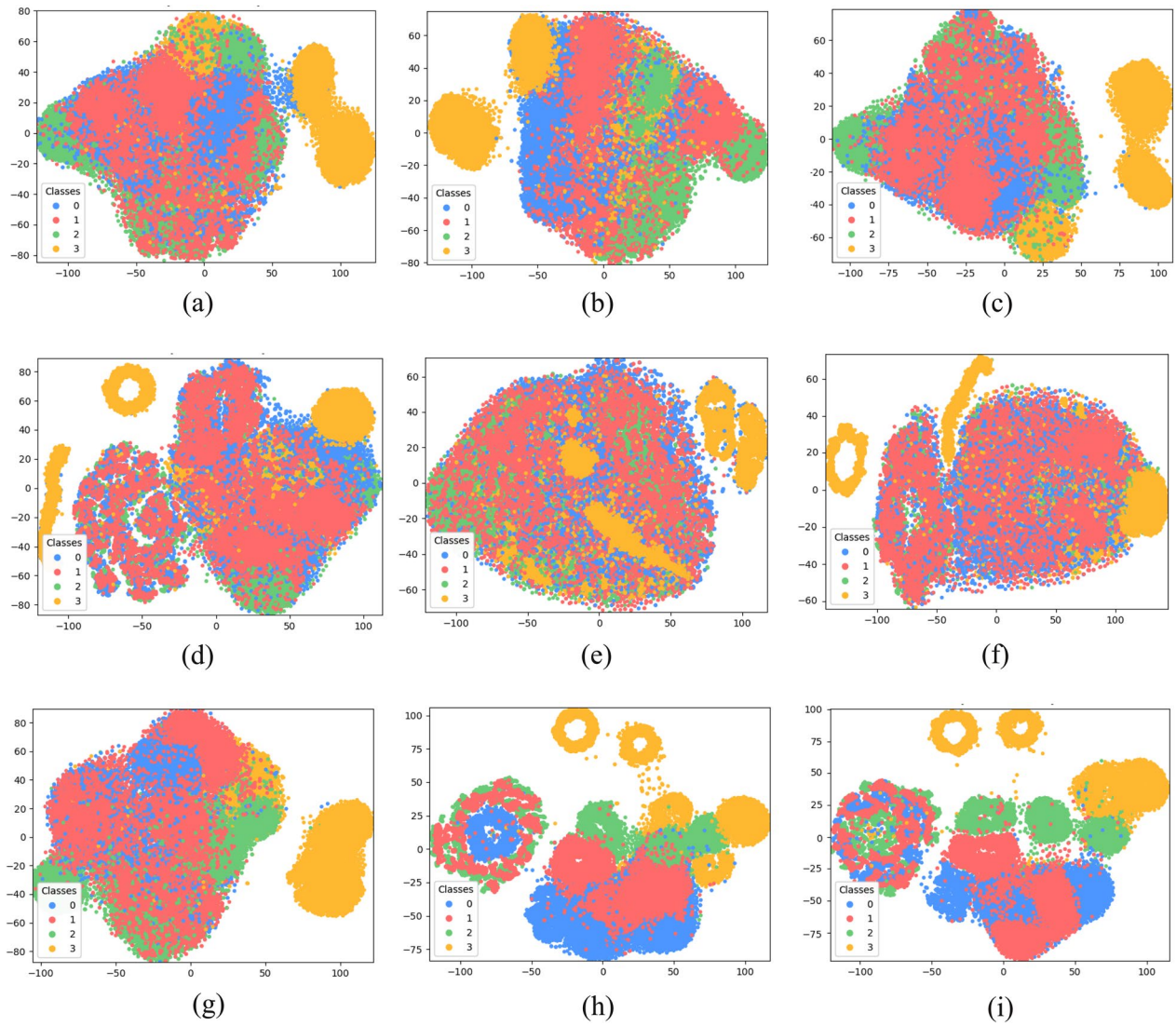

Figure 14 Visualization of features using t-SNE: **a** CNN; **b** WCNN; **c** MKCNN; **d** WaveletNet-L; **e** WaveletNet-m; **f** WaveletNet-M; **g** RSN; **h** SincNet; **i** DFAWNet

5 Conclusions

In this study, the SPINN, an intelligent diagnosis framework, is presented with effectively taking advantage of both SP-based methods and DL-based methods. For the

Table 5 Experimental results of different models

Model	Max-acc (%)	Min-acc (%)	Avg-acc (%)
CNN	76.42	74.67	75.55 ± 0.55
WCNN	89.52	88.77	88.47 ± 0.62
MKCNN	87.77	86.03	86.72 ± 0.49
DSN	83.84	80.79	82.27 ± 0.79
SincNet	90.39	89.52	89.74 ± 0.29
DFAWNet	92.36	93.32	92.76 ± 0.49

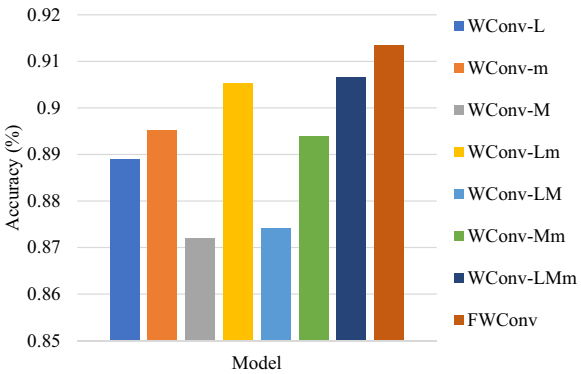

Figure 15 Diagnostic results of CNN with different first layers

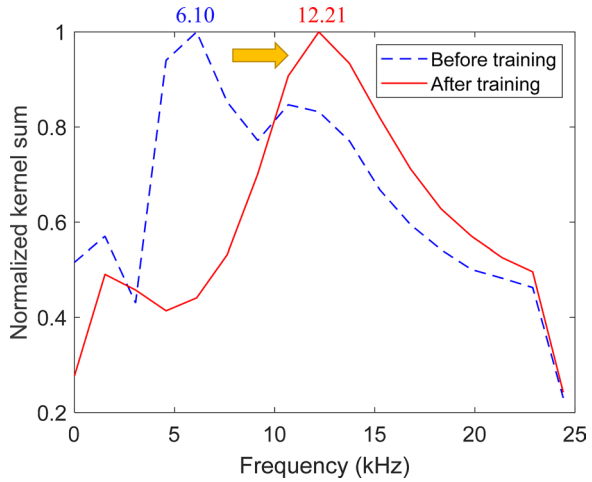

Figure 16 Cumulative frequency band of the wavelet convolution kernel

requirement of weak fault feature extraction under heavy noise, we design the SPINN with wavelet denoising and index-based filtering. As a DL-based implementation for SPINN, DFAWNet has explainability, denoising ability, fault feature extraction ability, and end-to-end parameter learning ability. The FWConv alleviates the requirement of wavelet basis and scale selection for wavelet transform implementation. DHT denoises signals with dynamic feature estimation in inter-scale and intra-scale. In addition, ISF optimizes and selects optimal wavelet filters for diagnostic feature extraction. As an integration, it's convenient to embed the DFAW module in front of a DL model and improve its performance.

Table 6 Experimental results of the ablation experiment

Model	Max-acc (%)	Min-acc (%)	Avg-acc (%)
DFAWNet	93.32	92.36	92.76 ± 0.49
DFAWNet-1	92.62	91.35	92.11 ± 0.48
DFAWNet-2	91.27	90.39	90.87 ± 0.39
DFAWNet-3	92.67	92.05	92.43 ± 0.25

Different experiments on three datasets verified the performance of the proposed DFAWNet. With added noise, the diagnostic results confirmed that DFAWNet has better robustness against noise than other methods. In the second experiment, we tested all methods on variable operating conditions. DFAWNet has the least amount of performance degradation. Moreover, the ablation experiment evaluated the effectiveness of each component. The visualization of the feature after thresholding and the cumulative frequency band of the wavelet kernel illustrate the explainability of this work.

However, the selection of the wavelet bases is still restricted by prepared base types and the fusion method is still designed based on priors. The exploration of extracted fault features is not enough. The decision stage of the SPINN is still not explainable. Further research might explore replacing handcrafted wavelet bases with some constraints related to the filter property. In addition, the correlation between the selected feature and the fault should be studied. For different application requirements, other designs to SPINN are also expected. Moreover, we believe that the SPINN is useful for analyzing

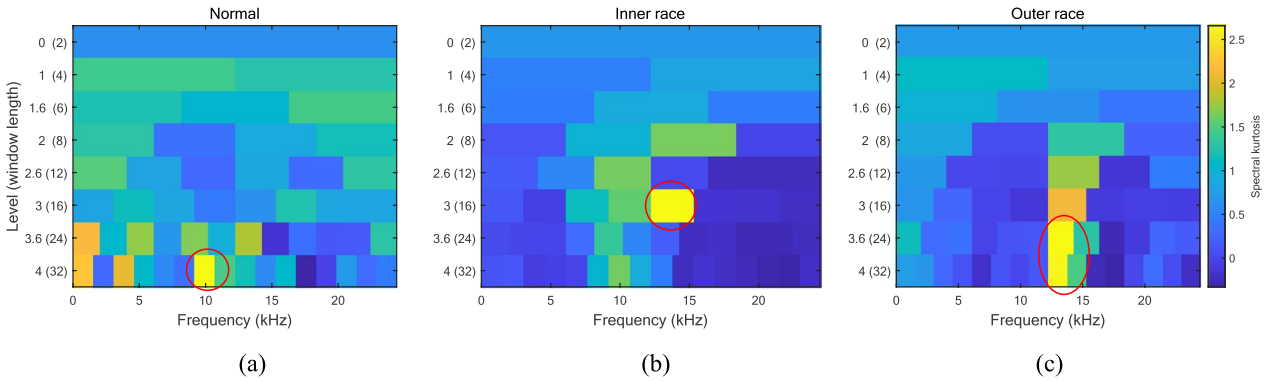

Figure 17 The paving spectral kurtosis values and their associated frequency bands for signals in different states: **a** normal state (the optimal central frequency is 9.92 kHz); **b** inner race fault (the optimal central frequency is 13.73 kHz); **c** outer race fault (the optimal central frequency is 13.22 kHz)

other non-vibration signals, such as audio signals or physiological signals.

Acknowledgements
The authors sincerely thank Zheng Zhou, Tianfu Li for their help on this work.

Author contributions
ZS: Writing-review & editing; ZZ: Writing-review & editing; RY: Review-editing & supervision. All authors read and approved the final manuscript.

Authors' Information
Zuogang Shang received the B.S. degree in mechanical engineering from *Xi'an Jiaotong University, China*, in 2020. He is currently working toward his Ph.D degree in mechanical engineering at *Xi'an Jiaotong University, China*. His current research is focused on explainable deep learning, mechanical fault diagnosis, and anomaly detection.
Zhibin Zhao received the B.S. and the Ph.D. degrees in mechanical engineering from *Xi'an Jiaotong University, China*, in 2015 and 2020, respectively. He is currently an Assistant Professor in mechanical engineering with *School of Mechanical Engineering, Xi'an Jiaotong University, China*. His research interests include sparse signal processing and machine learning algorithms for machinery health monitoring and healthcare.
Ruqiang Yan received the Ph.D. degree in mechanical engineering from the *University of Massachusetts at Amherst, MA, USA*, in 2007. He is currently a professor of mechanical engineering with *Xi'an Jiaotong University, China*. His research interests include data analytics, AI, and energy-efficient sensing for health diagnosis of large-scale, complex, dynamical systems. Dr. Yan is a fellow of *ASME* (2019) and *IEEE* (2022). He is also the Editor-in-Chief of *IEEE Transactions on Instrumentation and Measurement*, an Associate Editor of *IEEE Sensors Journal*, and Editorial Board Member of *Chinese Journal of Mechanical Engineering* and *Journal of University of Science and Technology of China.*

Funding
Supported by National Natural Science Foundation of China (Grant Nos. 51835009, 52105116) and China Postdoctoral Science Foundation (Grant Nos. 2021M692557, 2021TQ0263).

Competing interests
The authors declare no competing financial interests.

Received: 12 October 2022 Revised: 14 December 2022 Accepted: 5 January 2023
Published online: 23 January 2023

References

[1] Z Zhao, S Wu, B Qiao, et al. Enhanced sparse period-group lasso for bearing fault diagnosis. *IEEE Transactions on Industrial Electronics*, 2018, 66(3): 2143-2153.
[2] L Liao, W Jin, R Pavel. Enhanced restricted Boltzmann machine with prognosability regularization for prognostics and health assessment. *IEEE Transactions on Industrial Electronics*, 2016, 63(11): 7076-7083.
[3] X Chen, M Ma, Z Zhao, et al. Physics-informed deep neural network for bearing prognosis with multi-sensory signals. *Journal of Dynamics, Monitoring and Diagnostics*, 2022, 1(4): 200-207.
[4] W Zhao, C Zhang, S Wang, et al. Rolling bearing remaining useful life prediction based on wiener process. *Journal of Dynamics, Monitoring and Diagnostics*, 2022, 1(4): 229-236.
[5] H Chen, R Liu, Z Xie, et al. Majorities help minorities: Hierarchical structure guided transfer learning for few-shot fault recognition. *Pattern Recognition*, 2022, 123: 108383.
[6] R Liu, B Yang, E Zio, et al. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 2018, 108: 33-47.
[7] S Li, Y Xin, X Li, et al. A review on the signal processing methods of rotating machinery fault diagnosis. *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, 2019: 1559-1565.
[8] J Antoni. A critical overview of the "Filterbank-Feature-Decision" methodology in machine condition monitoring. *Acoustics Australia*, 2021, 49(2): 177-184.
[9] R Yan, R X Gao, X Chen. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Processing*, 2014, 96: 1-15.
[10] J Chen, Y Zi, Z He, et al. Compound faults detection of rotating machinery using improved adaptive redundant lifting multiwavelet. *Mechanical Systems and Signal Processing*, 2013, 38(1): 36-54.
[11] R Yan, M Shan, J Cui, et al. Mutual information-assisted wavelet function selection for enhanced rolling bearing fault diagnosis. *Shock and Vibration*, 2015: 1-9
[12] D Wang. Some further thoughts about spectral kurtosis, spectral L2/L1 norm, spectral smoothness index and spectral Gini index for characterizing repetitive transients. *Mechanical Systems and Signal Processing*, 2018, 108: 360-368.
[13] Y Lei, B Yang, X Jiang, et al. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 2020, 138: 106587.
[14] J Wu, Z Zhao, C Sun, et al. Few-shot transfer learning for intelligent fault diagnosis of machine. *Measurement*, 2020, 166: 108202.
[15] R Liu, F Wang, B Yang, et al. Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions. *IEEE Transactions on Industrial Informatics*, 2019, 16(6): 3797-3806.
[16] X Yu, Z Zhao, X Zhang, et al. Deep-learning-based open set fault diagnosis by extreme value theory. *IEEE Transactions on Industrial Informatics*, 2021, 18(1): 185-196.
[17] D Gunning, M Stefik, J Choi, et al. XAI—Explainable artificial intelligence. *Science Robotics*, 2019, 4(37): eaay7120.
[18] Z Ye, J Yu. Deep morphological convolutional network for feature learning of vibration signals and its applications to gearbox fault diagnosis. *Mechanical Systems and Signal Processing*, 2021, 161: 107984.
[19] T Li, Z Zhao, C Sun, et al. WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, 52(4): 2302-2312.
[20] J Yuan, S Cao, G Ren, et al. LW-Net: an interpretable network with smart lifting wavelet kernel for mechanical feature extraction and fault diagnosis. *Neural Computing and Applications*, 2022, 34(18): 15661-15672.
[21] D Wang, Y Chen, C Shen, et al. Fully interpretable neural network for locating resonance frequency bands for machine condition monitoring. *Mechanical Systems and Signal Processing*, 2022, 168: 108673.
[22] G Michau, G Frusque, O Fink. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences*, 2022, 119(8): e2106598119.
[23] M Zhao, S Zhong, X Fu, et al. Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics*, 2019, 16(7): 4681-4690.
[24] D L Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 1995, 41(3): 613-627.
[25] R Yan, R X Gao. Harmonic wavelet-based data filtering for enhanced machine defect identification. *Journal of Sound and Vibration*, 2010, 329(15): 3203-3217.
[26] D Wang, Z Peng, L Xi. The sum of weighted normalized square envelope: A unified framework for kurtosis, negative entropy, Gini index and smoothness index for machine health monitoring. *Mechanical Systems and Signal Processing*, 2020, 140: 106725.
[27] J Antoni, R B Randall. The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical Systems and Signal Processing*, 2006, 20(2): 308-331.
[28] B Ganguly, S Chaudhuri, S Biswas, et al. Wavelet kernel-based convolutional neural network for localization of partial discharge sources within a power apparatus. *IEEE Transactions on Industrial Informatics*, 2020, 17(3): 1831-1841.
[29] M Lin, Ji R, Chen B, et al. Training compact CNNs for image classification using dynamic-coded filter fusion. *arXiv preprint*, 2021, arXiv:2107.06916.
[30] J Chen, Y Zi, Z He, et al. Adaptive redundant multiwavelet denoising with improved neighboring coefficients for gearbox fault detection. *Mechanical Systems and Signal Processing*, 2013, 38(2): 549-568.
[31] E Jang, S Gu, B Poole. Categorical reparametrization with Gumble-Softmax. *International Conference on Learning Representations (ICLR 2017)*, OpenReview. net, 2017.

[32] Y Bengio, Y Lecun, G Hinton. Deep learning for AI. *Communications of the ACM*, 2021, 64(7): 58-65.
[33] Y LeCun, L Bottou, Y Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
[34] M Claesen, M B De. Hyperparameter search in machine learning. *Proc. of the 11th Metaheuristics International Conference*, 2015: 1–5.
[35] B Wang, Y Lei, N Li, et al. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 2018, 69(1): 401-412.
[36] Z Zhao, T Li, J Wu, et al. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Transactions*, 2020, 107: 224-255.
[37] W Zhang, G Peng, C Li, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 2017, 17(2): 425.
[38] M Ravanelli, Y Bengio. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018: 1021–1028.
[39] E Bechhoefer. Condition based maintenance fault database for testing diagnostics and prognostic algorithms. *MFPT Data*, 2013.