Mechanical Systems and Signal Processing 195 (2023) 110314

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Interpretable convolutional neural network with multilayer wavelet for Noise-Robust Machinery fault diagnosis

Huan Wang [a,b], Zhiliang Liu [a,*], Dandan Peng [c], Ming J. Zuo [a,d,e]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China
[c] Department of Mechanical Engineering, KU Leuven, Leuven 3000, Belgium
[d] Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta T6G 1H9, Canada
[e] Qingdao International Academician Park Research Institute, Qingdao 266041, China

ARTICLE INFO

Keywords:
Fault diagnosis
Wavelet transform
Convolutional neural network
Attention mechanism

ABSTRACT

Convolutional neural networks (CNNs) are being utilized for mechanical fault diagnosis, due to its excellent automatic discriminative feature learning ability. However, the poor interpretability and noise robustness of CNNs have plagued both academia and industry. Since traditional signal analysis technology has a sound theoretical basis and physical meaning, it motivates us to use signal processing theory to improve the interpretability and performance of the CNN algorithm. To this end, this paper proposes a multilayer wavelet attention convolutional neural network (MWA-CNN) for noise-robust machinery fault diagnosis. This framework aims to learn discriminative fault features from the wavelet domain, which allows the model to obtain better interpretability and superior performance than conventional time-domain-based CNNs. The proposed Discrete Wavelet Attention Layer (DWA-Layer) is used to map time domain signals to wavelet space, and obtain valuable information through the learnable convolutional layer. By alternately using DWA-Layer and convolutional layer for signal decomposition and feature learning, the proposed framework actually embeds a similar multi-resolution analysis algorithm in CNN. This helps integrate physics-based knowledge into the CNN. Finally, the frequency attention mechanism is proposed to enhance the ability of MWA-CNN to obtain fault-related features from different frequency components. Experiments on high-speed aeronautical bearing and motor bearing datasets prove that the proposed method has excellent fault diagnosis ability and noise robustness. The visual analysis of the attention mechanism contributes to the interpretability of CNN in the field of fault diagnosis.

1. Introduction

Modern mechanical systems generally have complex mechanical structures and work in harsh environments. Some key mechanical components (such as bearings and gears) are widely used in important mechanical systems, including wind turbines, high-speed trains, and aero engines [1]. However, during operation of mechanical equipment, the bearing inevitably appears fatigue degradation, cracks, deformation, spalling and other failures. If faulty parts are not repaired or replaced in time, it may cause serious damage to whole mechanical system. Therefore, in order to ensure normal operation of mechanical systems, it is very necessary to monitor health status

* Corresponding author.
E-mail address: zhiliang_liu@uestc.edu.cn (Z. Liu).

https://doi.org/10.1016/j.ymssp.2023.110314
Received 6 July 2022; Received in revised form 7 February 2023; Accepted 19 March 2023
Available online 29 March 2023

of key components of mechanical systems [2].

Machinery condition monitoring based on vibration signal analysis is the current mainstream technology. However, signals collected by the sensor contain a lot of noise, which brings challenges to signal analysis [3]. This is because mechanical equipment is a complex system composing of multiple components, and shocks and vibrations caused by other factors are transmitted among different components. In addition, mechanical systems also face interference from the outside world. In the initial stage of the fault, the fault characteristic information is weak, and it is easy to be overwhelmed by noise and difficult to detect [4]. To address these challenges, deep learning-based intelligent diagnostic techniques have received more and more attention from academia and industry. The advantage of deep learning is that it can learn a set of good fault-related features from big dataset and can automatically diagnose the health status of mechanical equipment. In recent years, a variety of deep learning algorithms have been extensively studied and applied to machinery fault diagnosis [5–8]. For instance, Liu et al. [5] proposed an improved autoencoder based on recurrent neural network (RNN) for bearing fault diagnosis. Zhou et al. [8] proposed an improved generative adversarial network to solve the problem of sample imbalance. In particular, the convolutional neural network (CNN) stands out among these algorithms and obtains the state-of-the-art performance in a variety of fault diagnosis tasks [9–12]. For instance, Liu et al. [13] proposed a lightweight multi-task CNN architecture for fault diagnosis and condition monitoring of wheelset bearings. Chen et al. [14] Combined CNN with extreme learning machine, and achieved good results on multiple datasets. Han et al. [15] proposed a CNN architecture for vibration signal denoising. This method has good denoising performance on bearing dataset and can improve the diagnostic performance of the CNN model in noisy environments.

Although CNN has achieved good performance in many tasks, it still has the following problems.

1) CNN mainly consists of finite filters, and they are learned by a stochastic gradient descent algorithm under random initial conditions. Compared with methods such as wavelet transform, CNN is quite rough in terms of signal processing. This rough method requires a large number of parameters to obtain sufficient feature expression abilities.
2) It is becoming more and more important in practical industrial applications to understand how an algorithm learns and what it learns. However, CNN has a large number of parameters that are difficult to analyze, so it is still used as a black-box model and difficult for CNN to perform an in-depth interpretability analysis.
3) In practical applications, noise is inevitable. According to our experiments and literature reports [15], the CNN algorithm is susceptible to noise, that is, noise seriously affects the performance of the CNN model. The worthless noise will conceal the valuable information of the signal, causing the algorithm to overfit the invalid features.

To enhance the interpretability of CNN in intelligent diagnosis, many efforts have been made. Wang et al. [16] used the attention method to explore the feature learning mechanism of CNN, and similarly, Yang et al. [17] combined gate recurrent unit (GRU) with the attention mechanism and then analyzed the interpretability of the neural network through attention. Zhou et al. [18] proposed a partially interpretable neural network for fault diagnosis of gas turbines. Since traditional signal analysis methods have a sound theoretical basis and clear physical meaning, we believe that using them to improve the interpretability of CNNs is a very feasible solution. Li et al. [19] designed a continuous wavelet convolutional layer to replace the first convolutional layer of the standard CNN, which enabled the CNN to gradually incorporate the advantages of wavelet transform and its good interpretability. Further, our study deeply explores the fusion architecture of CNN and wavelet transform and proposes a deep hybrid architecture of wavelet and convolution. The proposed method is no longer limited to the first layer of the network; a hybrid architecture of wavelet and convolutional layer runs through the entire neural network framework. Convolutional layer and wavelet transform are alternately performed layer by layer.

Based on the above discussion, this paper aims to rethink the feature learning of CNN architecture from the wavelet domain. Compared with the time domain, the frequency domain or wavelet domain has always been an excellent space for signal processing problems in the field of fault diagnosis, such as signal denoising and fault feature extraction [20]. Compared with randomly initialized convolution kernels, wavelet analysis has excellent feature learning efficiency and interpretability. Wavelet analysis is also a powerful tool for signal denoising. Therefore, we believe that integrating the discrete wavelet transform (DWT) in the whole network can not only improve the model's shortcomings in signal analysis and reduce the number of its parameters, but also make the model have the ability to filter noise and invalid information and have good noise robustness. In addition, the excellent theoretical foundation of wavelet analysis also lays a good foundation for the interpretability of the model.

Therefore, this paper attempts to expand the learning space from the time domain space to the more useful wavelet domain space, so as to break through the current bottleneck of CNN in performance and interpretability. To this end, this paper proposes a discrete wavelet attention layer (DWA-Layer). DWA-Layer extends the feature learning space to the wavelet domain space, which uses discrete wavelet transform to decompose the signal into multiple frequency components. The frequency attention mechanism (FAM) is proposed to further enhance the ability of DWA-Layer to learn valuable information and filter irrelevant noise. DWA-Layer can be embedded in any position of the CNN architecture to jointly participate in model training and gradient update. Inspired by wavelet packet analysis algorithms, CNN-based feature learning and DWT-based signal decomposition are performed alternately, and the fault-related features hidden in the time domain and wavelet domain are learned layer by layer. This is equivalent to embedding a similar wavelet packet decomposition algorithm in the CNN architecture, which is usually used for multi-resolution analysis. Finally, a novel CNN framework based on multilayer wavelet attention (MWA-CNN) for mechanical fault diagnosis is proposed. MWA-CNN is dedicated to learning discriminative fault features from the wavelet domain. It uses efficient wavelet transform to improve the deficiencies of CNN in signal processing; relies on wavelet analysis with domain knowledge to significantly enhance the interpretability of the algorithm; adaptively filters irrelevant information through frequency attention mechanism to retain valuable features.

Additionally, one research approach is to integrate CNN with an independent signal preprocessing step for fault diagnosis [21–24]. Although these methods can improve the performance of CNN to some extent, they cannot solve the shortcomings of deep models. The ideas of this study are fundamentally different from them. In this study, wavelet transform and CNN are deeply fused, and they are alternately integrated in the whole network. This means that the wavelet transform and convolutional layers jointly participate in model training and gradient update. The update of CNN parameters is affected by the result of wavelet transform, and the result of wavelet transform depends on the current parameters of CNN. They promote each other during model training and jointly learn valuable information from the signals.

The proposed method is verified on the high-speed aeronautical (HSA) bearing and motor bearing datasets, and the experimental results show that it significantly improves the diagnostic performance of the CNN model. The proposed MWA-CNN has good interpretability and has great practical application potential.

The contributions of this paper are summarized as follows:

1) This paper explores the expansion of the feature learning space of the diagnostic model to the wavelet domain space. The wavelet domain space can provide inherent advantages that the time domain space does not have, and brings new insights into the feature learning and interpretability of CNN.
2) This paper implements a wavelet-transform-based layer (DWA-Layer), which uses DWT to map the time domain space to the wavelet domain space, and proposes an attention mechanism to learn valuable frequency domain information. DWA-Layer can be embedded in any position of the deep learning model, and jointly participate in model training and gradient update.
3) This paper proposes a signal-processing-based CNN framework (MWA-CNN) that embeds DWT into deep learning model for machinery fault diagnosis. The framework deeply integrates DWT and convolutional layers, and the feature flow is processed and learned in the network model in a novel form.
4) This paper analyzes the interpretability of the proposed method, and the results show that it learns fault-related features layer by layer and suppresses irrelevant information.

The paper is organized as follows. Section II introduces the basic theory. Section III describes the DWA-Layer and MWA-CNN in detail. In Section IV, the effectiveness and superiority of DWA-Layer and MWA-CNN is verified. Section V discusses four aspects of MWA-CNN. Section VI analyzes the interpretability of the proposed method. Section VII summarizes this paper.

2. Basic theory

2.1. Wavelet transform

Let $\psi(t)$ be a square integrable function, that is,$\psi(t) \in L^2(R)$. A family of functions can be obtained from $\psi(t)$ after scaling and translation transformation.

$$\psi_{\tau,\upsilon}(t) = \frac{1}{\sqrt{\tau}}\psi(\frac{t-\upsilon}{\tau}),\ \tau,\upsilon \in R,\ \tau \neq 0 \tag{1}$$

$\{\psi_{\tau,\upsilon}\}$ is called continuous wavelet, ψ is basic wavelet or mother wavelet and has $\widehat{\psi}(t=0) = 0$. τ is the scaling factor, and υ is the translation factor. The scaling factor changes the shape of the continuous wavelet. The translation factor changes the displacement of the continuous wavelet.

For the initial signal $x(t) \in L^2(R)$, its continuous wavelet transform (CWT) is defined as:

$$W_x(\tau,\upsilon) = \langle x, \psi_{\tau,\upsilon}\rangle = \frac{1}{\sqrt{\tau}}\int_{-\infty}^{+\infty} x(t)\overline{\psi}(\frac{t-\upsilon}{\tau})dt \tag{2}$$

where $\overline{\psi}(t)$ represents the complex conjugate of $\psi(t)$. Both τ and υ are continuous variables. It can be seen that $W_x(\tau,\upsilon)$ represents the projection of the signal $x(t)$ on the wavelet basis function $\psi(t)$.

By discretizing the scaling factor τ and the translation factor υ, the DWT can be obtained. In general, τ and υ are defined as:

$$\tau = \tau_0^i,\ \upsilon = j\upsilon_0\tau_0^i \tag{3}$$

where i and j are integers. The discrete wavelet is defined as:

$$\psi_{\tau,\upsilon} = \frac{1}{\sqrt{\tau_0^i}}\psi(\frac{t-j\upsilon_0\tau_0^i}{\tau_0^i}) = \tau_0^{-i/2}\psi(\tau_0^{-i}t - j\upsilon_0) \tag{4}$$

The corresponding DWT is:

$$W_x(\tau_0^i, j\upsilon_0\tau_0^i) = \tau_0^{-i/2}\int_{-\infty}^{+\infty} x(t)\psi(\tau_0^{-i}t - j\upsilon_0)dt \tag{5}$$

2.2. Multi-Resolution analysis and wavelet packet analysis

Multi-resolution analysis was first proposed by Mallat [25] in 1989, and it provides a new way to implement wavelet analysis algorithms. Let $X(n)$ be the discrete sequence of signal $x(t)$, where $n = 1, 2, \ldots, N$. If the decomposition scale $\eta = 0$, the signal can be expressed as $X(n) = A_0(n)$. Then the decomposition algorithm can be expressed as:

$$\begin{cases} A_\eta(n) = \sum_{k\in Z} h(k-2n) A_{\eta-1}(k) \\ D_\eta(n) = \sum_{k\in Z} g(k-2n) A_{\eta-1}(k) \end{cases} \tag{6}$$

where $h(n)$ and $g(n)$ are the filter coefficients determined by the wavelet basis function. $h(n)$ and $g(n)$ have low-pass and high-pass properties, respectively. η is the number of decomposed layers. The wavelet decomposition tree for multi-resolution analysis is shown in Fig. 1(a).

The discrete sequence A_0 is decomposed into two signal components through a low-pass filter and a high-pass filter, and then sampled at intervals, and finally the low-frequency component A_1 and the high-frequency component D_1 of the signal are obtained. A_1 is selected to repeat the decomposition steps described above to complete the multi-resolution decomposition of the signal. It can be seen that multi-resolution analysis is actually equivalent to multiple band-pass filters. Only low-frequency components are decomposed in each layer, and high-frequency components are not considered. One advantage of wavelet packet analysis is that it decomposes the frequency of the signal in multiple levels. It not only decomposes the low-frequency components of each layer, but also decomposes the high-frequency components of each layer. The result of a 3-layer wavelet packet decomposition is shown in Fig. 1(b). It can be seen that D_1 is also decomposed into its low frequency component AD_2 and high frequency component DD_2. Then, AD_2 and DD_2 are further decomposed.

3. Multilayer wavelet attention CNN

3.1. Discrete wavelet attention Layer

DWT is an effective signal processing technique, which can decompose a signal into two wavelet coefficients, namely low-frequency component and high-frequency component. Further decomposing these two wavelet coefficients can display the information of different frequency bands of the signal. This paper proposes a novel discrete wavelet attention layer (DWA-Layer). DWA-Layer introduces advantages of wavelet transform to the CNN model, so that the improved CNN model can not only learn the information hidden in the time domain and frequency domain, but also has good interpretability. In order to facilitate understanding, we will first introduce the proposed discrete wavelet transform layer (DW-Layer), and then introduce its improved version: DWA-Layer. Finally, the wavelet-based error propagation is analyzed.

3.2. Discrete wavelet transform Layer

In the CNN model, the data stream is generally characterized in the form of feature maps. Assuming that $M \in \Re^{C\times L}$ is the feature map of one-dimensional CNN (1DCNN), then M is a two-dimensional matrix with length L and width C. C represents the number of channels, and L represents the length of feature signals. That is, for the feature map M, it contains C feature signals. Based on the given wavelet basis function, a low-pass filter $h(n)$ and a high-pass filter $g(n)$ can be obtained, and $g(k) = (-1)^k h(1-k)$. It can be seen from Eq. (8) that after each signal is filtered by $h(n)$ and $g(n)$, its low-frequency and high-frequency components can be finally obtained. In 1DCNN model, DW-Layer decomposes each feature signal independently. Assuming that the feature map M is the input of DW-Layer, the output of DW-Layer is the low-frequency feature map $LM \in \Re^{C\times L'}$ and the high-frequency feature map $HM \in \Re^{C\times L'}$ respectively. Generally, $L' = L/2$, and the value of L' varies slightly when different wavelet basis functions are used. Subsequently, LM and HM are spliced into a whole according to the channel to obtain the mixed feature $LG \in \Re^{2C\times L'}$. Compared with learning from the feature map M, the CNN model can learn richer feature information from the wavelet domain feature map.

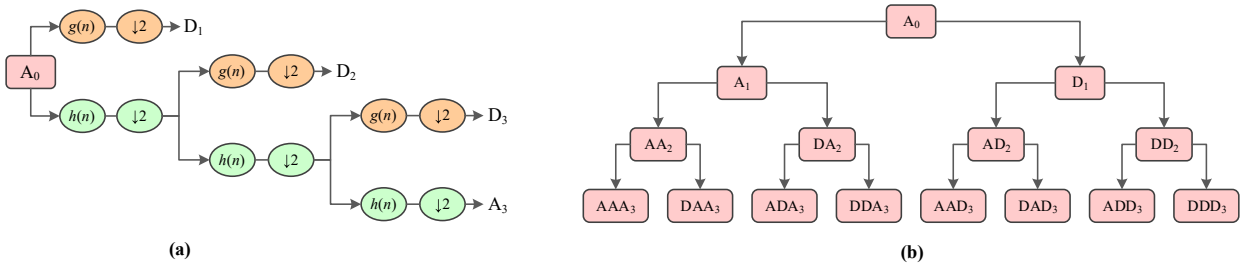


Fig. 1. The decomposition structure of multi-resolution analysis (a) and wavelet packet analysis (b).

Dropout, as an effective regularization method to suppress over-fitting, is usually used in the fully connected layer. We found that dropout technology also has excellent anti-overfitting ability in DW-Layer. Because the existence of DW-layer makes it easier for CNN model to learn features from the signal, the overfitting phenomenon of CNN model becomes serious. Dropout technology can solve this problem well. Therefore, after each DW-Layer, Dropout is used to further improve the generalization performance of the CNN model.

Although DW-Layer can decompose the input feature signal, filtering out the fault-related feature from the obtained results is still a complex problem. In conventional signal analysis methods, the recognition of features can be achieved through some indicators, but this is difficult to achieve in the CNN architecture. Therefore, this paper proposes DWA-Layer to make the CNN model automatically pay attention to important feature information and reduce the model's attention to irrelevant information (noise, etc.). DWA-Layer introduces the frequency attention mechanism (FAM) to aggregate the global information of the mixed feature *LG*, and then encodes the relative importance of different channel features to guide the feature learning of the CNN model.

3.3. Frequency attention mechanism

As shown in Fig. 2, DWA-Layer is mainly composed of a DW-Layer and a FAM. The core idea of FAM is to use the self-learning ability of CNN to filter out the signal components that are useful for the diagnosis task from the mixed feature $LG \in \Re^{2C\times L'}$. It replaces manual feature selection with an automatic valuable feature attention mechanism similar to human visual attention. First, FAM introduces a global average pooling layer (GAP) [26] to compress the global information of the mixed feature *LG* into a channel representation vector $z \in \Re^{2C\times 1}$. The *i*-th element of **z** is expressed as:

$$z_i = \mathrm{GAP}(LG_i) = \frac{1}{L'}\sum_{j}^{L'} LG_i(j) \tag{7}$$

LG_i represents the *i*-th feature signal of *LG*. Then, FAM uses a simple encoding and decoding mechanism to capture the importance of these channel signals. The channel representation vector is first compressed into a hidden layer vector with a dimension of $C \times 1$, and then decoded back to the original dimension. Finally, the channel weight vector $z' \in \Re^{2C\times 1}$ is output. The encoding and decoding operations of FAM are completed by two convolutional layers respectively. The first convolutional layer uses the ReLU function to provide non-linear transformation capabilities. The second convolutional layer uses the Sigmoid function, which is mainly used to map the obtained feature vector to an interval range of 0 to 1, thereby generating a weight vector z'. The size of the element of z' represents the importance of the corresponding channel feature. This can be expressed as:

$$\begin{aligned} \hat{z} &= \mathrm{ReLU}(\boldsymbol{W}_1 * z + b_1),\ \hat{z} \in \Re^{C\times 1}, \\ z' &= \mathrm{Sigmoid}(\boldsymbol{W}_2 * \hat{z} + b_2),\ z' \in \Re^{2C\times 1}. \end{aligned} \tag{8}$$

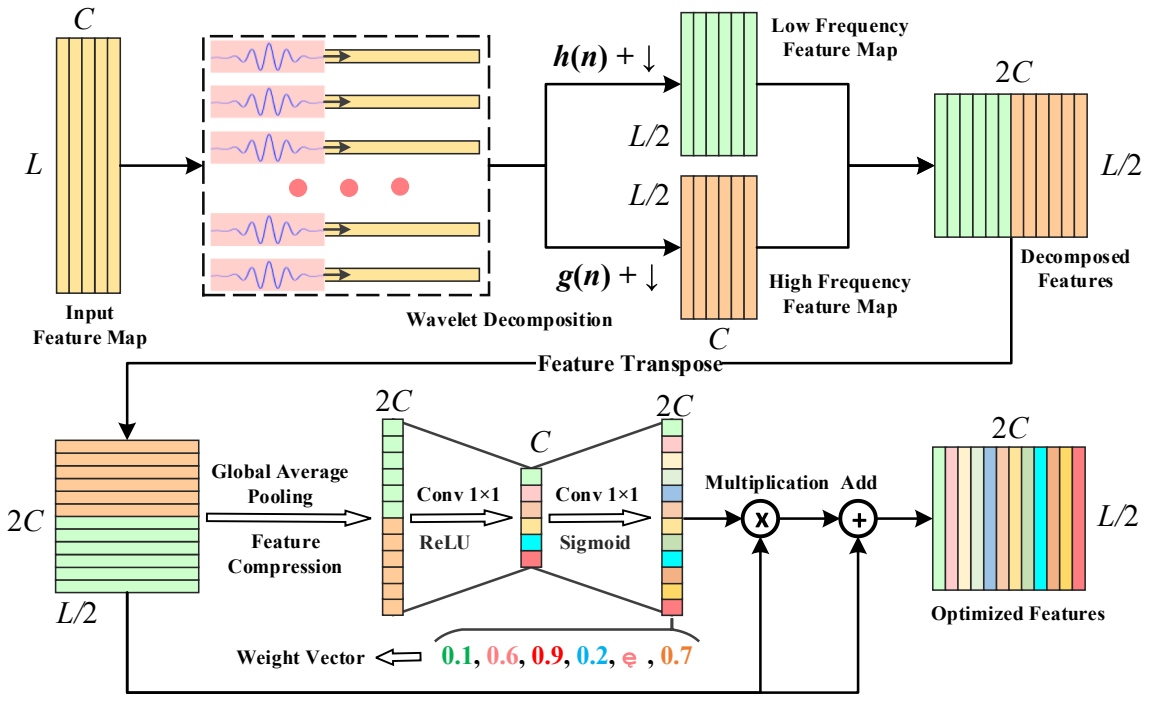


Fig. 2. The architecture details of the proposed DWA-Layer, which mainly consists of a DW-layer and a frequency attention mechanism.

Finally, FAM uses matrix multiplication to embed this weight information into the CNN model to guide the feature learning of the CNN model. In addition, residual connections are added in FAM to optimize the gradient propagation of the network and prevent feature responses from being too small.

3.4. Wavelet-Based backpropagation

This work deeply integrates DWT and convolution to jointly participate in optimization and gradient propagation. Wavelet-driven backpropagation is described below. Assume the error passed to the output of LG is $\xi_{LG}^{\varphi} \in \mathbb{R}^{2C\times L'_2}$, and the low-pass and high-pass filters for the DWT are h^{η},g^{η}, φ denotes the φ_{th} layer. Since in the forward operation, the output of the DWT is concatenated according to their channels, then the backpropagation firstly needs to spilt ξ_{LG}^{φ} into two parts:ξ_{LM}^{φ}, $\xi_{HM}^{\varphi} \in \mathbb{R}^{C\times L'_2}$. ξ_{LM}^{φ} and ξ_{HM}^{φ} correspond to the low-frequency and high-frequency output of the DWT. Note that the split is needed to take the same sequence as the forward concatenation.

Before error is back-propagated, ξ_{LM}^{φ} and ξ_{HM}^{φ} are firstly up-sampled by a factor of two, since the forward computation is performed by a convolution with a step of two:

$$\widetilde{\xi}_{LM,HM}^{\varphi}(n) = \begin{cases} \xi_{LM,HM}^{\varphi}(n/2), & n/2 \text{ are integers} \\ 0, & otherwise \end{cases} \tag{9}$$

Then error is passed to the input of DWT by:

$$\xi_{\Im}^{\varphi-1} = \widetilde{\xi}_{LM}^{\varphi} * reverse(h^{\eta}) + \widetilde{\xi}_{HM}^{\varphi} * reverse(g^{\eta}) \tag{10}$$

3.5. Multilayer wavelet attention convolutional neural network

The CNN model has good automatic feature learning ability, and DWT has good signal decomposition ability. The effectiveness of these two technologies in fault diagnosis tasks has been verified by numerous studies. This paper attempts to combine the advantages of these two technologies to propose a novel machinery fault diagnosis model. Fig. 3 shows the proposed Multilayer Wavelet Attention CNN. MWA-CNN is mainly composed of multiple convolutional layers and DWA-Layer. The collected signal is first decomposed by

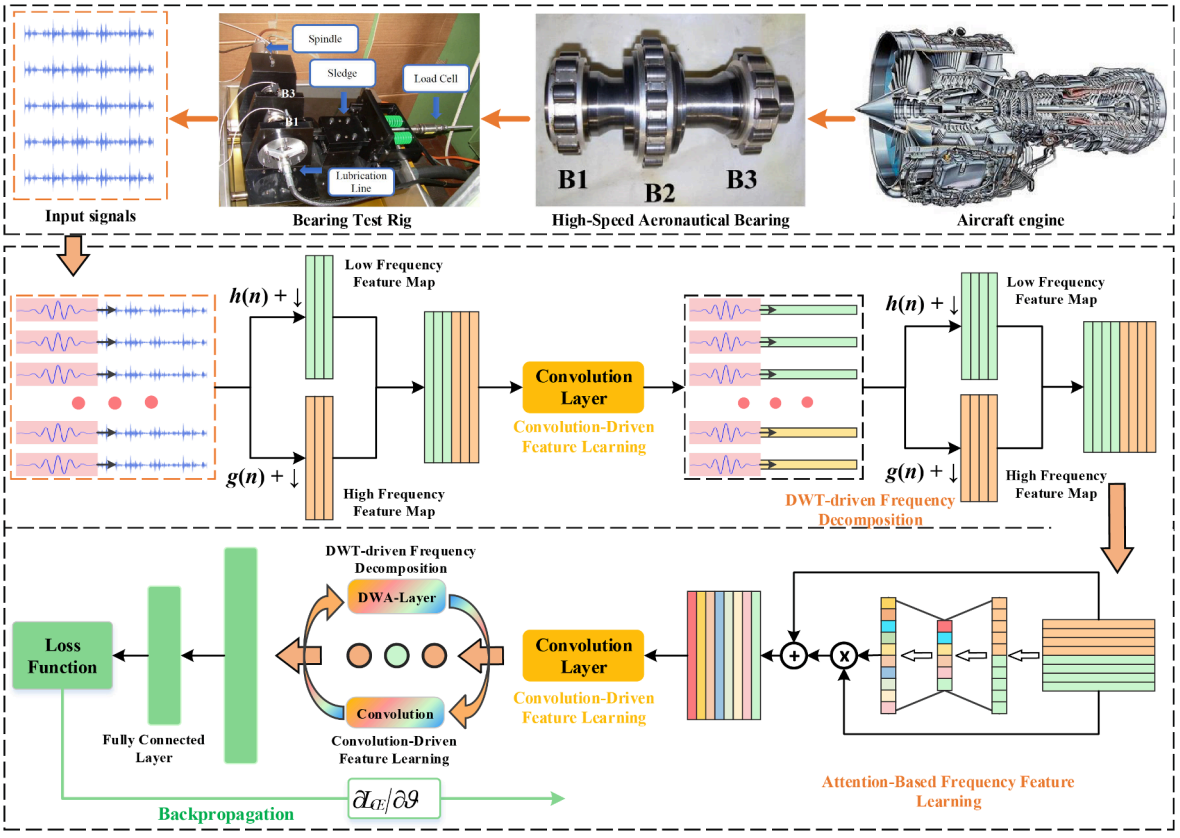


Fig. 3. The detailed network architecture of the proposed multilayer wavelet attention CNN (MWA-CNN).

DWA-Layer, and then input into the convolutional layer for automatic feature learning. In the entire MWA-CNN, the above operations are performed multiple times. This means that the input signal is decomposed layer by layer into multiple frequency components, and the required feature information is also learned layer by layer by the convolutional layer. In addition, this paper recommends using the proposed DWA-Layer to replace MaxPooling. DWA-Layer performs DWT decomposition of the signal while performing feature downsampling, so as to display the hidden features in the time domain and frequency domain to the convolutional layer. This paper uses the group normalization (GN) layer [27] to replace the batch normalization (BN) layer. The GN layer divides the channels into groups and calculates the mean and variance within each group for normalization. Although the performance of GN is similar to that of BN when using a large batch size, the performance of GN is significantly better than BN when using a small batch size. In fault diagnosis, obtaining enough labeled samples to train the model in many scenarios is difficult. In this scenario, GN will perform better than BN. Therefore, adopting GN can give the proposed method broad applicability. The training algorithm of MWA-CNN is shown in **Algorithm. 1**. $D = \left\{\Gamma^i, j^i\right\}_{i=1}^{N}$ represents the dataset, Γ^i represents the data sample, and its corresponding label is.

Generally, the fault diagnosis task can be regarded as a multi-classification task. Therefore, Softmax is adopted, which transforms the input features into a probability distribution whose sum is 1. Assuming that the output of the fully connected layer is a logit vector ϖ, χ represents the input signal of the CNN model, and the total number of fault categories is S.

$$p(s|\chi) = \frac{\exp(\varpi_s)}{\sum_{s=1}^{S}\exp(\varpi_s)} \tag{11}$$

$p(s|\chi)$ is the predicted probability that χ belongs to category s. Suppose $q(s|\chi)$ is the true probability that χ belongs to category s. The cross-entropy loss function is used to measure the distance between the prediction and the true label, which can be expressed as:

$$L_{CE} = -\sum_{s=1}^{S} q(s|\chi)\log p(s|\chi). \tag{12}$$

Algorithm 1 Multilayer Wavelet Attention Convolutional Neural Network

Input: Dataset $D = \left\{\Gamma^i, j^i\right\}_{i=1}^{N}$ trained by mini-batch
Output: Optimized Predictor $F(v)$
1: Set $e = 0$ and epoch E; set H.
2: Initialize the ϑ; Normalize the dataset D
3: **while** $e \le E$ **do**
4: Sample a batch of data B from D; set $\hbar = 1$.
5: Calculate DWT according to Eq. (8), obtained A_1 and D_1, then splice them.
6: Perform CNN to extract features:$F_1^C = \mathrm{Re}LU\left(W^T(A_1 + D_1) + b\right)$
7: **while** $\hbar \le H$ **do**
8: Calculate DWT according to Eq. (8), obtained A_h and D_h
9: Get mixed frequency features:$LG = (A_h + D_h)$
10: Perform attention to filter frequency features:$z^{'} \otimes LG + LG$
11: Perform CNN to extract features:$F_h^C = \mathrm{Re}LU\left(W^T(z^{'} \otimes LG + LG) + b\right)$
12: Set $\hbar = \hbar + 1$
13: Calculate the output of
14: Calculate, update by back propagation
15: If an epoch is completed, set $e = e + 1$
16: **end while**

Table 1
The Parameter Configuration of MWA-CNN.

Layer	Type	Kernel	Channel	Output
1	Input Layer	N/A	N/A	4096 × 1
2	DWA-Layer	N/A	N/A	2048 × 2
3	CNN-Layer	3 × 1	12	2048 × 12
4	DWA-Layer	N/A	N/A	1024 × 24
5	CNN-Layer	3 × 1	24	1024 × 24
6	DWA-Layer	N/A	N/A	512 × 48
7	CNN-Layer	3 × 1	48	512 × 48
8	DWA-Layer	N/A	N/A	256 × 96
9	CNN-Layer	3 × 1	96	256 × 96
10	DWA-Layer	N/A	N/A	128 × 192
11	CNN-Layer	3 × 1	192	128 × 192
12	DWA-Layer	N/A	N/A	64 × 384
13	CNN-Layer	3 × 1	384	64 × 384
14	Global Average Pooling			384
15	FC + Softmax			7

Table 1 shows the parameter configuration of MWA-CNN. When MWA-CNN is applied to the HSA bearing dataset, the input dimension of MWA-CNN is 4096 × 1. DWA-Layer is used to replace the MaxPooling layer to complete feature dimensionality reduction. Not only that, DWA-Layer can also perform multi-resolution signal decomposition on the input signal. The output dimension of each layer varies slightly depending on the wavelet basis function and the padding. Assuming the Haar wavelet is used, the output dimension of the DWA-Layer is half of the input dimension. MWA-CNN consists of six CNN-Layers and six DWA-Layer. The number of CNN-Layers and DWA-Layer can be simply adjusted to suit different tasks. A CNN-Layer contains a convolutional layer, a GN layer, and a ReLU activation function. The kernel of the convolutional layer is set to 3 × 1, the number of channels gradually increases from 12 to 384, and the padding is set to 0. The classification layer of MWA-CNN consists of a GAP and a fully connected layer with a Softmax function. The open-source code of the proposed method can be found at ***https://github.com/PHM-Code/MWA-CNN***.

3.6. Architecture design combining wavelet and CNN

In this study, we deeply integrate wavelet transform and CNN, so that they can learn from each other and jointly promote each other in a unified model. In the proposed MWA-CNN, convolutional layers and wavelet transform are performed alternately to jointly build a deep neural architecture. This actually builds a deep frequency feature decomposition and feature learning mechanism. With the deepening of the network depth, the frequency features are gradually decomposed in detail, and valuable features are also learned layer by layer by the convolution layer.

To highlight the characteristics of the proposed method, Fig. 4 shows different architectural design methods combining wavelet transform and CNN. **(1) Ordinary CNN Network.** It follows the end-to-end design concept, and its network architecture is stacked by multiple convolutional layers. The approach does not use signal analysis methods. **(2) CNN with Signal Analysis** [21–24]. It tried to combine CNN with signal analysis methods such as wavelet transform or Fourier transform. This method first uses signal analysis techniques to preprocess the signal and then uses the CNN model for feature learning and fault identification. This method only uses wavelet transform as a data preprocessing method, and CNN and wavelet transform are not perfectly combined. Most existing methods combining wavelets and CNNs can be classified into this category [28,29]. **(3) Wavelet-Kernel-Net** [19]. It designs a continuous wavelet convolution (CWConv) layer to replace the first convolutional layer of standard CNN. This enables the first CWConv layer to discover more meaningful kernels. This method fully integrates CNN and wavelet transform, but is limited to the first layer of the model. **(4) Our Proposed Method.** It deeply integrates wavelet transform and convolutional layers, which enables the proposed method to perfectly fuse the advantages of both to achieve effective fine-grained signal decomposition and fault-related feature learning. Moreover, an attention mechanism is proposed to enable the model to distinguish valuable frequency features and ignore irrelevant information.

As shown in Fig. 5, from the perspective of signal analysis, our architecture design draws on the ideas of multi-resolution analysis and wavelet packet decomposition. Moreover, unlike wavelet packet decomposition, MWA-CNN has self-learning ability. The convolutional layer is placed after the signal decomposition layer to learn fault-related features. The convolutional layer is actually used as a feature selector, which sends the useful feature signals of the previous layer to the next layer (DWA-Layer) for more detailed decomposition. As the number of layers increases, the signal is decomposed more finely by DWA-Layer, and the convolutional layer can learn the feature information hidden in the time domain and frequency domain well. Through layer-by-layer signal decomposition and feature learning, irrelevant information (noise, etc.) is gradually removed, so that discriminative fault features can be learned by MWA-CNN.

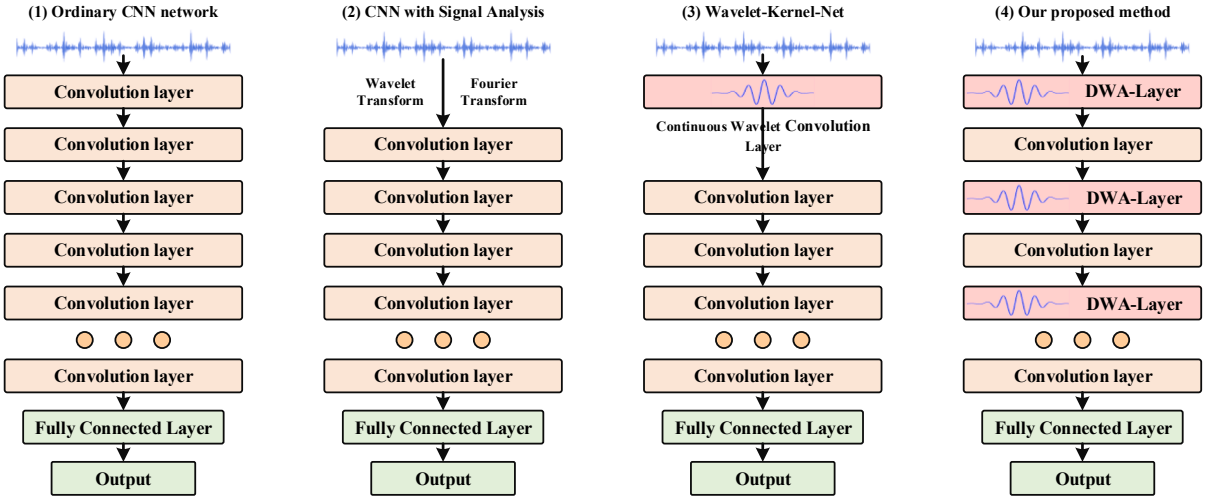


Fig. 4. The architectural design methods combining wavelet transform and CNN.

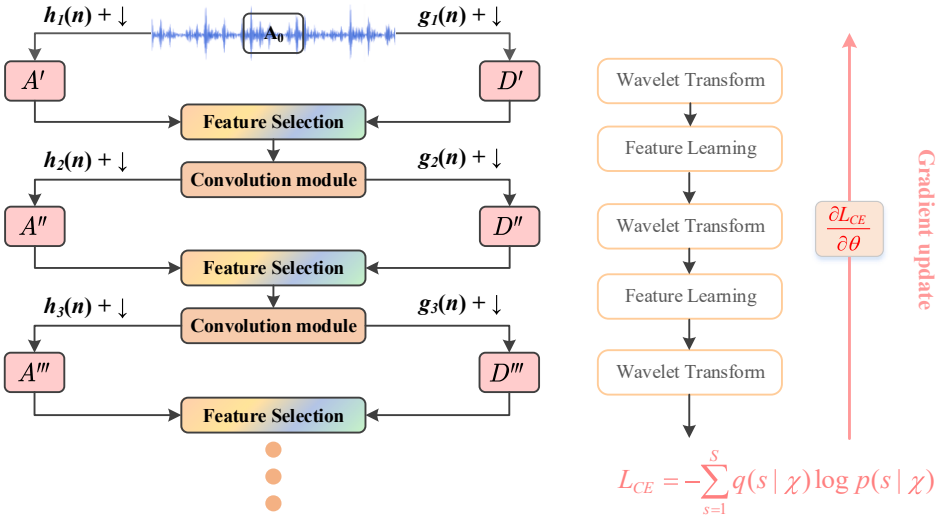


Fig. 5. This illustrates the architectural design idea of the proposed method.

4. Experimental validation

In this section, the effectiveness of the proposed DWA-Layer and MWA-CNN is verified on the HSA bearing dataset and the motor bearing dataset. In addition, MWA-CNN is compared with cutting-edge deep learning methods. We also proved that MWA-CNN has good noise robustness.

4.1. Experimental setup

The proposed method is implemented by deep learning framework Pytorch, python 3.7 and pytorch_wavelets [30]. All models are trained and tested on a server with NVIDIA GeForce RTX 3090 GPU. The server has 32 GB memory and uses Intel 10900 K CPU. Z-score normalization is used to normalize all data samples so that the training of the model becomes stable. In the training process, the Adam optimization algorithm is adopted, which has the advantages of fast calculation efficiency and small memory requirements. Adam can also accelerate the convergence speed of the network model. The batch size is set to 64 and the learning rate is set to 0.0001. Since Daubechies (db) wavelet is widely used in fault diagnosis tasks, this paper takes db16 wavelet as an example for experiment. It is worth

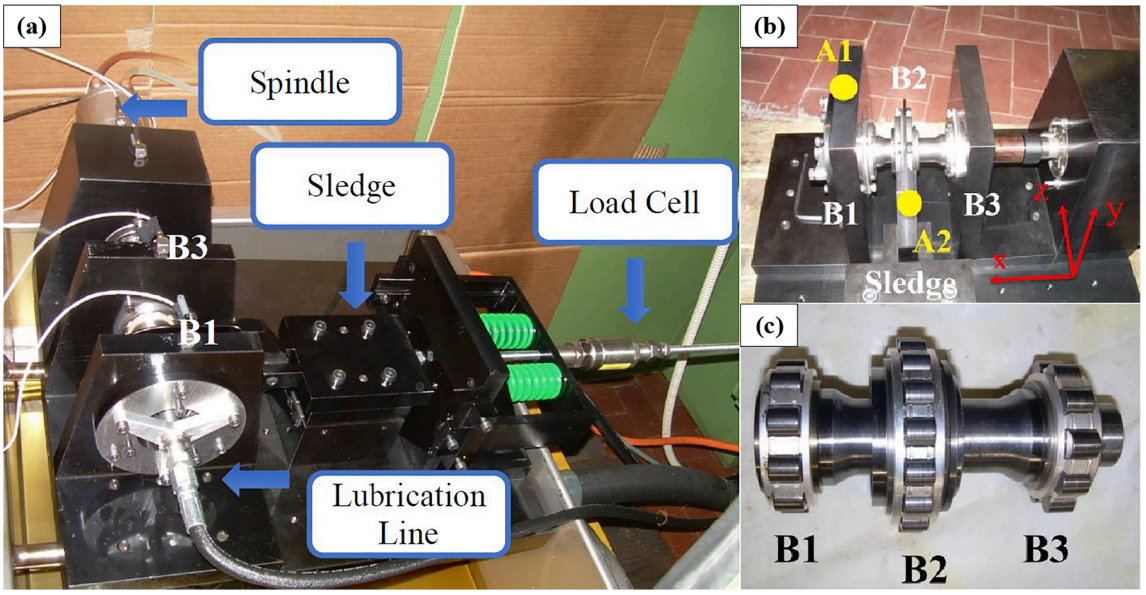


Fig. 6. The HSA bearings test rig a) general view of the test rig; b) positions of the accelerometers and the reference system; c) the shaft with its three roller bearings.

noting that the developed method is also applicable to other wavelets, such as Haar, Biorthogonal, Dmeyer, etc. This is discussed in Section V. This paper uses accuracy to evaluate the performance of the network model. Each experiment was repeated four times, and the mean and standard deviation were used as the final experimental results.

The collected vibration signal already contains a certain degree of noise. However, in real situations, the vibration signal may contain higher levels of noise. In order to verify the performance of MWA-CNN in different noise environments, we add additional Gaussian white noise to the vibration signal. In this paper, the Gaussian white noise with SNR = 4 dB, 0 dB, −2dB, and − 4 dB is added to the signal, respectively.

B. Case 1. ***C. High-Speed Aeronautical Bearings Fault Diagnosis.***

4.2. Data Description

The experimental data comes from the HSA bearing signal acquisition test rig [31]. Fig. 6 shows the structure of the test rig, the test bearing and the location of the acceleration sensor. The test rig is mainly composed of a high-speed spindle, which is used to drive the rotation of the shaft. As shown in Fig. 6(b), a triaxal IEPE accelerometer is installed at A1 and A2, and the sampling frequency is set to 51200 Hz. As shown in Fig. 6(c), the inner rings of these bearings (B1, B2, and B3) are connected to a very short and tick hollow shaft, specifically designed for speeds up to 35,000 rpm. As shown in Table 2, when collecting signals, seven health states are set on the B1 bearing. A total of one health state and six failure states. There are mainly two types of faults: inner ring fault and roller fault. These two kinds of faults have three different fault sizes, and their fault diameters are 150, 250 and 450 μm respectively. The experiment used HSA bearings, so the experiment was carried out under different loads and speeds. See Table 2 for details, where 100 Hz means 6000 rpm. In this experiment, we use the sliding segmentation method for data enhancement. The length of the signal sample is set to 4096 × 1, and a total of 22,134 training samples and 7259 test samples are obtained.

4.3. The effectiveness of DWA-Layer

This experiment compared four different network architectures. 1DCNN: this architecture only uses six convolutional layers and a classification layer with GAP. The parameter configuration of the convolutional layer and classification layer is the same as that of MWA-CNN. W-CNN: this architecture adds a DW-Layer before 1DCNN. MW-CNN: the parameter configuration of this architecture is the same as that of MWA-CNN, but all attention modules are removed. MWA-CNN: the method described in Section III. In order to illustrate the performance of these methods under different noise conditions, the experiment was carried out under four noise conditions (4 dB, 0 dB, −2dB and − 4 dB). Table 3 shows their experimental results.

As shown in Table 3, 1**DCNN** obtains an accuracy of 91.16% when SNR = 4 dB, however, when SNR = − 4 dB, it only obtains an accuracy of 50.38%. This shows that the ordinary CNN model can obtain high accuracy when there is no noise or less noise, but when the noise is strong, the performance of the ordinary CNN model is greatly reduced. This phenomenon has been reported by many studies. Therefore, improving the anti-noise ability of the CNN model is also the focus of this paper. After adding a DW-Layer before 1DCNN, we found that the performance of the model is slightly improved under all noise conditions. This shows that adding DWT is helpful to improve the performance of the CNN model. Furthermore, we add multiple DW-layers to the 1DCNN model, which is MW-CNN. The diagnostic performance of MW-CNN has been greatly improved compared to 1DCNN and W-CNN. For example, when SNR = 4 dB, MW-CNN obtains a diagnosis accuracy of 95.08%, which is 3.92% higher than 1DCNN. When SNR = − 4 dB, the diagnostic accuracy of MW-CNN is 70.66%, which is 20.28% higher than 1DCNN. This shows that the proposed method is effective, it integrates the advantages of CNN method and DWT technology, so that the network model has very good noise robustness. Then, we use the self-learning ability of the CNN model and introduce the attention mechanism to make the network automatically select useful frequency component information. Finally, MWA-CNN is proposed. MWA-CNN demonstrates excellent fault diagnosis performance. For example, when SNR = 4 dB, MWA-CNN obtains an accuracy of 98.75%. When SNR = − 4 dB, MWA-CNN still achieves 87.61% accuracy, which is 16.95% higher than MW-CNN. This shows that using attention to make the CNN model focus on useful feature information and ignore irrelevant information (such as noise) can greatly improve the anti-interference ability of the CNN model.

Fig. 7 and Fig. 8 show the training curves of 1DCNN, MW-CNN and MWA-CNN when SNR = 4 dB and SNR = − 4 dB. When SNR = 4 dB, the training accuracy and validation accuracy of these three network models increase rapidly, and then tend to be stable when epoch greater than 80. However, the validation accuracy of MW-CNN and 1DCNN is significantly lower than the training accuracy, indicating that there is a serious overfitting. When SNR = − 4 dB, the convergence speed of MW-CNN and 1DCNN is obviously slower,

Table 2
Description of the HSA Bearing Dataset Information.

Defect	Dimension	Load	Speed	Label
No defect	–	0 N-1800 N	100 Hz-500 Hz	F1
Diameter of an indentation on the inner ring	450 μm	0 N-1800 N	100 Hz-500 Hz	F2
Diameter of an indentation on the inner ring	250 μm	0 N-1800 N	100 Hz-500 Hz	F3
Diameter of an indentation on the inner ring	150 μm	0 N-1800 N	100 Hz-500 Hz	F4
Diameter of an indentation on a roller	450 μm	0 N-1800 N	100 Hz-500 Hz	F5
Diameter of an indentation on a roller	250 μm	0 N-1800 N	100 Hz-500 Hz	F6
Diameter of an indentation on a roller	150 μm	0 N-1800 N	100 Hz-500 Hz	F7

Table 3
The Experimental Results of 1DCNN, W-CNN, MW-CNN, and MWA-CNN Under Four Kinds of Noise.

Noise	1DCNN	W-CNN	MW-CNN	MWA-CNN
4 dB	91.16 ± 0.32	91.41 ± 0.96	95.08 ± 0.46	**98.75 ± 0.20**
0 dB	78.54 ± 0.29	79.28 ± 0.48	88.28 ± 0.83	**96.48 ± 0.20**
−2 dB	65.04 ± 1.61	68.45 ± 0.73	80.66 ± 0.33	**93.62 ± 0.38**
−4 dB	50.38 ± 0.64	54.01 ± 1.31	70.66 ± 1.10	**87.61 ± 0.85**

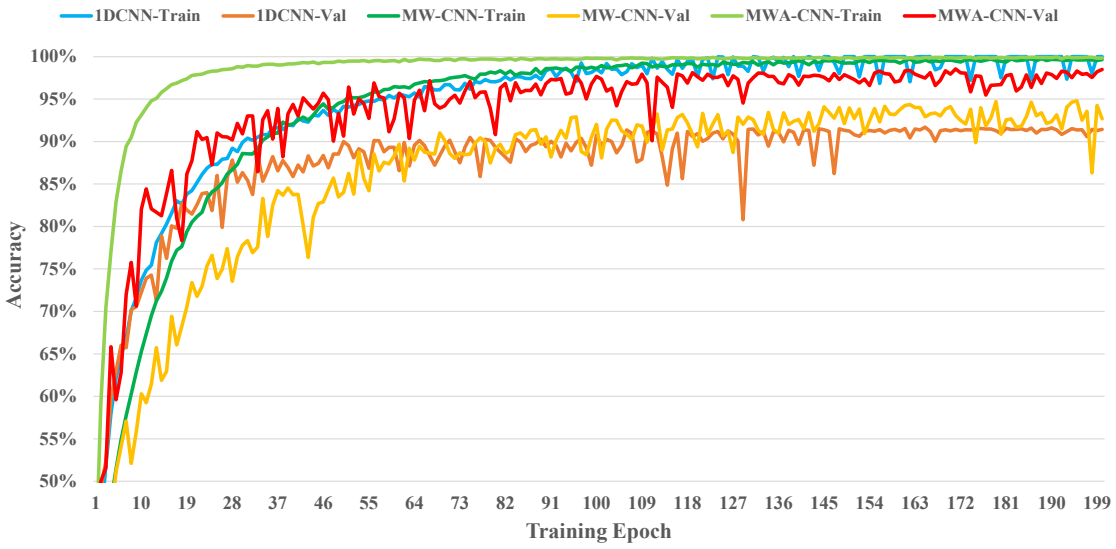


Fig. 7. The training accuracy curves and validation accuracy curves of 1DCNN, MW-CNN and MWA-CNN on the HSA bearing dataset (SNR = 4 dB).

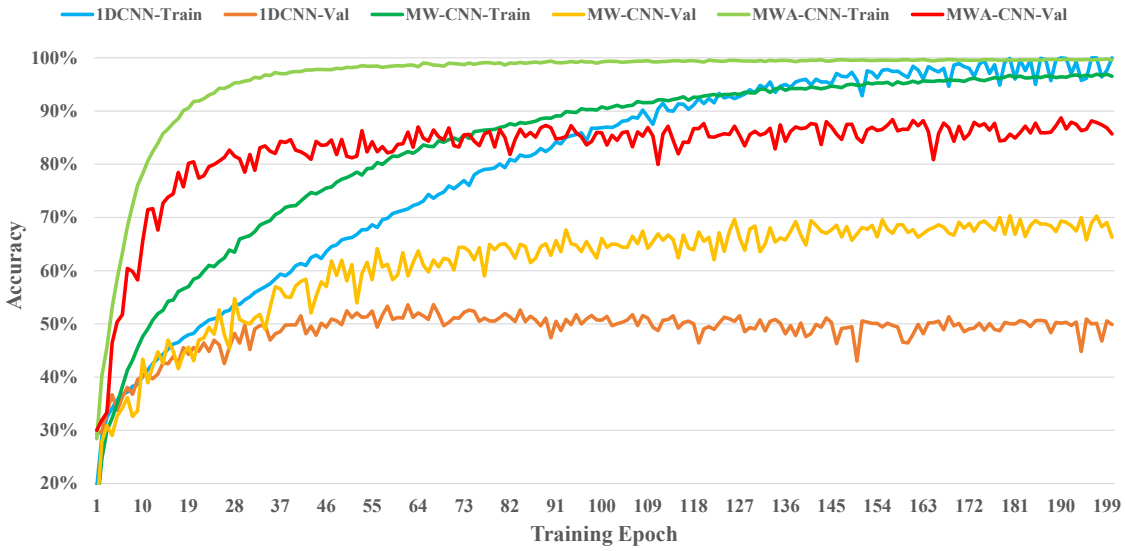


Fig. 8. The training accuracy curves and validation accuracy curves of 1DCNN, MW-CNN and MWA-CNN on the HSA bearing dataset (SNR = −4 dB).

which indicates that it is a great challenge to learn useful features from vibration signals in strong noise environment. The over-fitting phenomenon of MW-CNN and 1DCNN has also become more serious. When epoch = 180, the training accuracy of these two networks can reach more than 95%, and the validation accuracy is about 70% and 50%, respectively. In addition, MWA-CNN also has over-fitting phenomenon, but its situation is much better than MW-CNN and 1DCNN. Its validation accuracy can reach more than 85%.

In addition, we also show the visualization of the features of the fully connected layers of 1DCNN, MW-CNN and MWA-CNN in 2D

space. The visualization result is shown in Fig. 9, which is realized by T-SNE technology. Different colors represent different categories. When SNR = 4 dB, the features of MWA-CNN have very good distinguishability, and different categories are clearly distinguished. The features of 1DCNN and MW-CNN have poor distinguishability. When SNR = − 4 dB, the features of 1DCNN are completely indistinguishable, and samples of different fault categories are mixed together. This leads to poor fault diagnosis performance of the network. This also shows that it is difficult for 1DCNN to learn useful information from the signal under a strong noisy environment. MW-CNN performs better than 1DCNN, but its distinguishability is also poor. The features of MWA-CNN also have good distinguishability, indicating that the attention mechanism can effectively help MWA-CNN learn useful features from the signal.

4.4. Performance comparison

This experiment compares five existing deep learning methods to verify the superiority of the proposed method under noise conditions. These methods are MA1DCNN, WenCNN, ResNet-18, VGG-16 and LSTM. MA1DCNN was proposed by Wang et al. [16] for fault diagnosis of wheelset bearings. MA1DCNN is composed of multiple joint attention modules and convolutional layers, which can optimize the extracted features. WenCNN was proposed by Wen et al. [32] for fault diagnosis of motor bearings. WenCNN converts the time domain signal into a 2D image, and then uses a 2D CNN to extract feature information from the image. ResNet-18 [33] and VGG-16 [34] are two excellent network architectures for image recognition. We modify their 2D convolution into 1D convolution for machinery fault diagnosis. LSTM is an excellent network model, which has excellent long-term feature learning capabilities. These methods all adopt the same training strategy, and the experimental results under four kinds of noise are shown in Table 4.

As Table 4, MWA-CNN has the best performance under the four noise conditions. For example, when SNR = 4 dB, the diagnostic accuracy of MWA-CNN is 98.75%, which is 6.69% higher than MA1DCNN. When SNR = -4dB, the diagnostic accuracy of MWA-CNN is 87.61%, which is 22.8% higher than MA1DCNN. Thanks to the help of the attention mechanism, MA1DCNN has achieved the best performance among the five comparison methods, but there is still a big gap compared to MWA-CNN. In addition, these five comparison methods are very susceptible to noise, resulting in a significant performance degradation. For example, when SNR = 4 dB, MA1DCNN can obtain 94.06% accuracy, but when SNR = − 4 dB, its accuracy is only 64.81%. When SNR = 4 dB, VGG-16 can obtain 91.55% accuracy, but when SNR = − 4 dB, its accuracy is only 48.61%. In particular, when the SNR changes from 4 dB to 0 dB, the accuracy of these five comparison methods all drop by more than 10%. This shows that the anti-noise performance of these methods is insufficient, and it is difficult for them to obtain enough useful feature information from the noisy signal. MWA-CNN uses DWA-Layer to decompose the signal, making it more convenient to obtain useful fault features from the signal, and then obtain excellent diagnostic performance.

Fig. 10 and Fig. 11 show the confusion matrix of MWA-CNN when SNR = 4 dB and SNR = − 4 dB, respectively. The value on the diagonal indicates the number of samples that are correctly classified for each category. The last column indicates the number of test samples for each category. When the noise intensity is small (SNR = 4 dB), good diagnosis results can be obtained for most categories. The diagnosis recall of F1, F2, F3, F6 and F7 are all above 98%, and some are even close to 100%. When the SNR changes from 4 dB to − 4 dB, the recall and precision of each category are greatly reduced. For example, the recall of F1 dropped from 98.94% to 89.30%. The precision of F1 dropped from 99.03% to 88.70%. In particular, the recall of F4 dropped from 94.41% to 81.00%. This shows that this category is easily affected by noise. This may be because the fault diameter of F4 is only 150 μm, which is smaller than that of F2

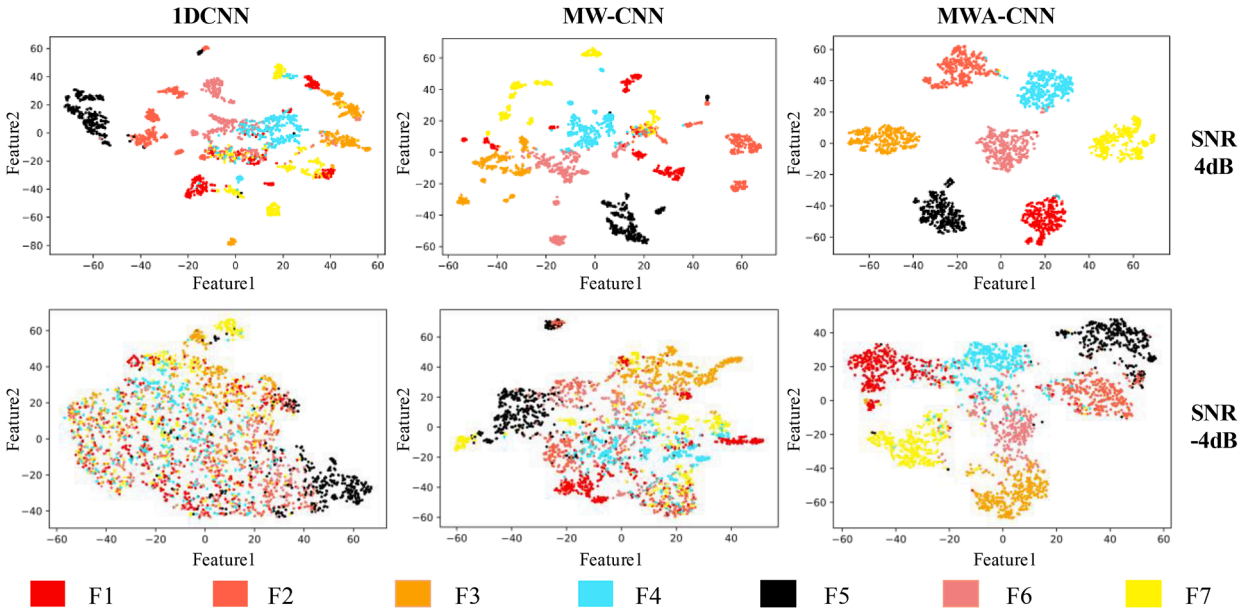


Fig. 9. The visualization of the features of the fully connected layer of 1DCNN, MW-CNN and MWA-CNN in 2D space.

Table 4
The Experimental Results of MWA-CNN and Five Comparison Methods Under Four Kinds of Noise.

Noise	MWA-CNN	MA1DCNN	WenCNN	ResNet-18	VGG-16	LSTM
4 dB	**98.75 ± 0.20**	94.06 ± 0.77	90.44 ± 0.96	85.31 ± 0.67	91.55 ± 0.48	87.23 ± 0.23
0 dB	**96.48 ± 0.20**	84.11 ± 0.90	78.91 ± 0.76	63.58 ± 0.56	79.18 ± 0.42	74.59 ± 1.02
−2 dB	**93.62 ± 0.38**	76.72 ± 0.89	72.10 ± 0.54	48.06 ± 1.03	65.50 ± 0.82	63.43 ± 2.28
−4 dB	**87.61 ± 0.85**	64.81 ± 1.13	61.18 ± 2.06	38.50 ± 0.87	48.61 ± 1.45	57.28 ± 0.93

True Label \ Predicted Label	F1	F2	F3	F4	F5	F6	F7	Recall	Test Number
F1	1026	0	0	8	0	3	0	**98.94%**	1037
F2	2	1028	0	5	0	1	1	**99.13%**	1037
F3	1	0	1033	0	0	3	0	**99.61%**	1037
F4	7	26	1	979	0	24	0	**94.41%**	1037
F5	0	0	0	23	1014	0	0	**97.78%**	1037
F6	0	0	7	9	0	1021	0	**98.46%**	1037
F7	0	0	0	1	0	0	1036	**99.90%**	1037
Precision	**99.03%**	**97.53%**	**99.23%**	**95.51%**	**100%**	**97.05%**	**99.90%**	—	7259

Fig. 10. The Confusion matrix of MWA-CNN on HSA bearing dataset (SNR = 4 dB).

True Label \ Predicted Label	F1	F2	F3	F4	F5	F6	F7	Recall	Test Number
F1	926	26	30	28	1	15	11	**89.30%**	1037
F2	21	921	10	40	12	10	23	**88.81%**	1037
F3	1	0	1000	2	0	14	20	**96.43%**	1037
F4	47	42	5	840	3	65	35	**81.00%**	1037
F5	4	29	0	26	975	1	2	**94.02%**	1037
F6	37	20	103	99	3	737	38	**71.07%**	1037
F7	8	12	6	11	7	25	968	**93.35%**	1037
Precision	**88.70%**	**88.71%**	**86.66%**	**80.31%**	**97.40%**	**85.01%**	**88.24%**	—	7259

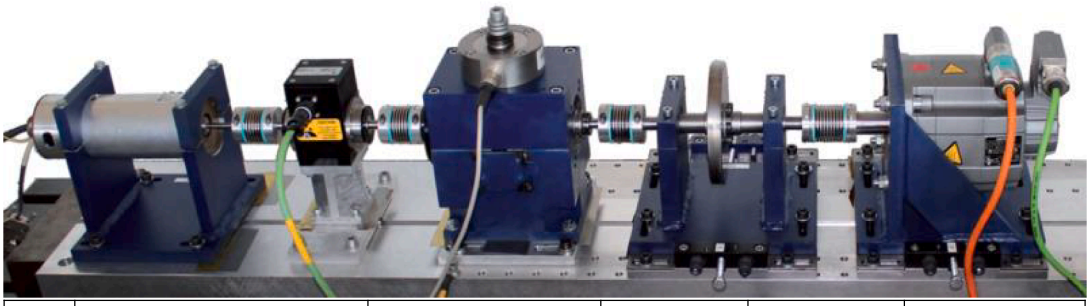
Fig. 11. The Confusion matrix of MWA-CNN on HSA bearing dataset (SNR = − 4 dB).

and F3. This leads to weaker fault features, which is easily overwhelmed by noise. In addition, category F6 is also easily affected by noise, which causes the model to incorrectly predict it into categories F3, F4, F7, etc.

G. Case 2. *H. **Motor Bearing Fault Diagnosis.***

4.5. Data Description

This dataset comes from the motor bearing signal acquisition experiment platform [35]. The test rig used is shown in Fig. 12. In addition, Fig. 12 also shows the fault information and working status information of the experimental bearing. The test rig consists of several modules: an electric motor, a torque-measurement shaft, a rolling bearing test module, a flywheel and a load motor.



Label	Damage	Bearing Element	Radial Force	Load Torque	Rotational Speed
H1	Normal	–	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm
H2	Fatigue: Pitting	Outer Ring	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm
H3	Plastic Deform.: Indentations	Outer Ring	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm
H4	Fatigue: Pitting	Outer Ring; Inner Ring	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm
H5	Plastic Deform.: Indentations	Outer Ring; Inner Ring	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm
H6	Fatigue: Pitting	Inner Ring	400N,1000N	0.1Nm,0.7Nm	900rpm,1500rpm

Fig. 12. The motor bearing test rig and the fault information of experimental bearings.

Experimental bearings with different failure modes are installed in the bearing test module to obtain experimental data. The failures of the experimental bearings are real and not artificial. The fault location is mainly divided into the inner ring and outer ring, and the failure mode is mainly fatigue pitting and plastic deform. As shown in Fig. 12, the dataset is divided into five different fault categories, plus the normal category, the dataset has a total of six categories. This experiment only uses vibration signal data, and its sampling frequency is 64 kHz. The sliding segmentation method is also used to increase the signal samples, and the length of each sample is 5120 $\times$ 1. A total of 36,053 training samples and 12,017 test samples were obtained.

4.6. *Performance comparison*

This section conducts experiments on MWA-CNN and five existing methods on the motor bearing dataset to show the performance of these methods on different datasets. The parameter configuration of these methods has been introduced in Section IV.B. Similarly, we also conducted experiments on these methods under 4 dB, 0 dB, −2dB and − 4 dB noise. The results are shown in Table 5.

Similarly, MWA-CNN performs better than five comparison methods under four kinds of noise conditions. When SNR = 4 dB, MWA-CNN achieved a diagnostic accuracy of 99.70%. In particular, when SNR = − 4 dB, MWA-CNN still achieved 98.71% accuracy. When the noise changes from 4 dB to − 4 dB, the accuracy of MWA-CNN only drops by less than 1%. This shows that MWA-CNN has very good noise robustness on this dataset, and it can obtain fault features from noisy signals very well. It can be found that the motor bearing dataset is relatively simple, and the performance of all methods is higher than that on the HSA bearing dataset. When SNR = − 4 dB, the accuracy of these five comparison methods are all over 85%. However, the performance of these methods has dropped significantly as the noise increases. In addition, different methods have different performance on different datasets. For example, VGG-16 does not perform well on the HSA bearing dataset, but obtains good results on the motor bearing dataset. MWA-CNN has obtained excellent results on these two datasets, which shows that MWA-CNN has good adaptability.

Fig. 13 and Fig. 14 show the confusion matrix of MWA-CNN when SNR = 4 dB and SNR = − 4 dB. On the motor bearing dataset, MWA-CNN can achieve good performance under these two noise conditions. In particular, when SNR = 4 dB, the precision of categories H1, H2 and H4 are all 100%. When the SNR changed from 4 dB to − 4 dB, only the Recall of category H6 showed a significant drop, and its Recall changed from 99.70% to 91.55%. The performance of other categories has not changed much. This shows that MWA-CNN has a very good performance on this dataset, and its performance does not drop significantly in a strong noise environment.

Table 5
The Experimental Results of MWA-CNN and Five Comparison Methods Under Four Kinds of Noise.

Noise	MWA-CNN	MA1DCNN	WenCNN	ResNet-18	VGG-16	LSTM
4 dB	**99.70 ± 0.03**	99.33 ± 0.38	95.07 ± 0.63	99.29 ± 0.21	99.65 ± 0.13	97.93 ± 0.25
0 dB	**99.65 ± 0.11**	98.17 ± 0.27	90.79 ± 1.10	97.91 ± 0.35	99.47 ± 0.20	95.98 ± 0.70
−2 dB	**98.99 ± 0.14**	89.16 ± 0.79	88.57 ± 0.38	95.58 ± 0.89	98.16 ± 0.47	94.22 ± 0.55
−4 dB	**98.71 ± 0.21**	86.58 ± 2.44	85.47 ± 0.94	91.62 ± 0.99	96.50 ± 0.24	92.97 ± 0.88

True Label \ Predicted Label	H1	H2	H3	H4	H5	H6	Recall	Test Number
H1	1994	0	0	0	7	0	**99.65%**	2001
H2	0	2979	0	0	4	29	**98.90%**	3012
H3	0	0	1999	0	0	0	**100%**	1999
H4	0	0	2	2004	0	0	**99.90%**	2006
H5	0	0	0	0	999	0	**100%**	999
H6	0	0	0	0	6	1994	**99.70%**	2000
Precision	**100%**	**100%**	**99.90%**	**100%**	**98.33%**	**98.57%**	—	12017

Fig. 13. The Confusion matrix of MWA-CNN on motor bearing dataset (SNR = 4 dB).

True Label \ Predicted Label	H1	H2	H3	H4	H5	H6	Recall	Test Number
H1	1994	3	0	0	4	0	**99.65%**	2001
H2	0	2988	0	2	1	21	**99.20%**	3012
H3	0	0	1999	0	0	0	**100%**	1999
H4	0	0	0	2006	0	0	**100%**	2006
H5	7	0	0	0	992	0	**99.30%**	999
H6	32	100	0	0	37	1831	**91.55%**	2000
Precision	**98.08%**	**96.67%**	**100%**	**99.90%**	**95.94%**	**98.97%**	—	12017

Fig. 14. The Confusion matrix of MWA-CNN on motor bearing dataset (SNR = − 4 dB).

4.7. Computational burden analysis

This section analyzes the computational burden of the proposed MWA-CNN and six other network models. Table 6 shows the parameters and Multiply-Accumulate Operations (MACs) of these methods. 1DCNN has the same network architecture as MWA-CNN but does not have the DWA-Layer. In other words, 1DCNN does not include the wavelet and attention mechanism. It can be seen that adding wavelet transform and frequency attention mechanism to the CNN model greatly increases the computational burden of the model. For example, the MACs of 1DCNN are only 0.02 G, and the MACs of MWA-CNN are 0.17 G. Although the computational burden has increased significantly, the benefits of adding DWA-Layers are also apparent. For example, on the HSA dataset with 0 dB noise, the performance of MWA-CNN is nearly 18% higher than that of 1DCNN. Moreover, the MACs of ResNet-18 and VGG-16 reach 0.44 G and 0.62 G, respectively. Compared with ResNet-18 and VGG-16, the computational burden of MWA-CNN is acceptable. The diagnostic performance of MWA-CNN is also significantly better than ResNet-18 and VGG-16. These results show that the proposed DWA-Layer can significantly improve the model's diagnostic performance, and the computational burden increase is acceptable.

Table 6
The Parameters and Computational Burden of MWA-CNN and the Six Comparison Methods.

-	MWA-CNN	1DCNN	MA1DCNN	WenCNN	ResNet-18	VGG-16	LSTM
Parameters	1.39 M	3.01 M	0.29 M	0.90 M	3.85 M	50.97 M	1.38 M
MACs	0.17 G	0.02 G	0.07 G	0.08 G	0.44 G	0.62 G	0.11 G

5. Discussions

5.1. Discuss the number of DWA-Layer

To demonstrate the effect of the number of DWA-Layer on model performance, we conduct experiments on the HSA bearing dataset with 4 dB noise. Seven different network architectures are experimented, namely MWA-CNN-2, MWA-CNN-3,…, MWA-CNN-8, where 2,3,…,8 represent the number of DWA-Layer. The experimental results are shown in Fig. 15. As the number of DWA-Layer increases, the performance of the diagnostic model also increases gradually. For example, the performance of MWA-CNN-2 is below 80%, while the performance of MWA-CNN-5 exceeds 95%. This shows that with the increase of DWA-Layer, the signal is decomposed more finely, and more valuable features are learned. When the number of DWA-Layer exceeds six, the performance of the diagnostic model does not increase but slightly decreases, which indicates that the model is over-fitting and no additional DWA-Layer is needed. Therefore, in this study, the number of DWA-Layer is set to six. In addition, for more complex and challenging tasks, the number of DWA-Layer can be increased to obtain better learning ability and diagnostic performance.

5.2. Comparing interpretable Space-Based approaches

The proposed method automatically obtains valuable features from the signal through a deep fusion architecture of wavelet transform and CNN. In addition, CNN can also learn valuable features from other interpretable feature spaces to complete fault prediction. For example, we can perform empirical mode decomposition (EMD) on the signal to get multiple intrinsic mode functions (IMFs) or perform fast Fourier transform (FFT) on the signal to get the frequency domain distribution. In addition, Wavelet-Kernel-Net combines continuous wavelet transform with the first convolutional layer to obtain valuable time–frequency information. To this end, this experiment conducts experimental analysis on MWA-CNN, EMD-CNN, FFT-CNN, Wavelet-Kernel-Net (Laplace-ResNet) [19], and 1DCNN, respectively. 1DCNN is the baseline model of the above network, mainly composed of six 1D convolutional layers. For EMD-CNN, the signal is first decomposed by EMD, and then 1DCNN is used to learn from multiple IMFs to complete fault prediction. For FFT-CNN, the signal is first decomposed by FFT, and then 1DCNN is used to learn from time and frequency domains to complete fault prediction. The backbone network of Laplace-ResNet is the standard ResNet-18 and uses the Laplace wavelet. The experiment used the motor bearing dataset and the results are shown in Table 7.

The results show that adopting an interpretable feature space can improve the performance of diagnostic models. This confirms that the feature space constructed by EMD, FFT, or wavelet is beneficial for CNN to learn valuable fault-related features. For example, compared with 1DCNN, the diagnostic performance of EMD-CNN and FFT-CNN is improved by 5.86% and 5.01% under −4 dB noise, respectively. The diagnostic performance of Laplace-ResNet is comparable to that of EMD-CNN. Through the deep integration of wavelet and CNN, the proposed MWA-CNN achieves the best performance, showing that it fully exploits the advantages of CNN and wavelet.

5.3. The influence of wavelet basis functions

In the experiment of Section IV, we only used the DB wavelet. In order to verify that the proposed method can also be applied to other wavelet basis functions, we further discussed the performance of other six wavelet basis functions on the HSA bearing dataset. The six wavelet basis functions are Haar, Dmeyer (Dmey), Symlet (sym8), Biorthogonal (bior3.1), Reverse Biorthogonal (rbio3.1) and Coiflet (coif8). The network architecture adopts MWA-CNN, and their experimental results under four kinds of noise conditions are shown in Fig. 16.

Obviously, using different wavelet basis functions, MWA-CNN shows different fault diagnosis performance. We found that db16 wavelet, sym8 wavelet, rbio3.1 wavelet and coif8 wavelet have relatively better performance on the HSA bearing dataset. For example, when SNR $= -4$ dB, the four wavelets mentioned above can achieve more than 85% diagnostic accuracy, while the accuracy of the other three wavelets is less than 85%. Although the diagnostic performance of the model changes when different wavelets are used, the magnitude of this change is not large. In general, these methods show good fault diagnosis ability. For example, when SNR $= 4$ dB, these methods have achieved a diagnostic accuracy of more than 97%. As the noise intensity increases, the performance gap among these wavelets gradually becomes obvious. When SNR $= -4$db, the worst-performing haar wavelet and bior3.1 wavelet still get more than 80% accuracy. Through the above experimental analysis, we suggest to adopt those wavelet basis functions with excellent performance, such as db16 wavelet, coif8 wavelet and rbio3.1 wavelet. Different wavelets have different characteristics. In future work, a multi-wavelet fusion strategy can be considered to make the model perform better.

5.4. The influence of the DB vanishing moments

In this section, we take dbN wavelet as an example to discuss the effect of the value of vanishing moments on network performance. We set a total of 18 network models, they use MWA-CNN and dbN wavelet, N is set to 2, 4, 6, 8, 10,…, 34, 36. The dataset uses the HSA bearing dataset. The experimental results under four different noise conditions are shown in Fig. 17.

Obviously, if different N is set for db wavelet, the model shows different diagnostic performance. When the noise is weak, these methods have good performance. For example, when SNR $= 4$ dB, the diagnostic accuracy of all models are higher than 97%. However, in the case of strong noise, and when N is less than 8, the model shows poor performance. For example, when SNR $= -4$ dB, the diagnostic accuracy of db2 is only 82%, which is significantly lower than the 87.3% accuracy of db10. When N is greater than 8, the

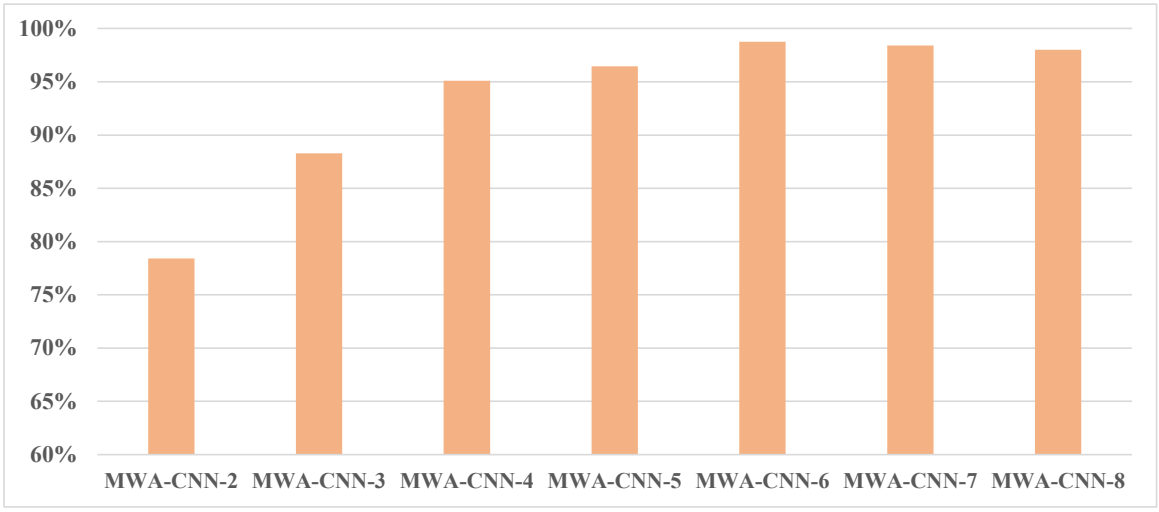


Fig. 15. Experimental results of the effect of the number of DWA-Layer on model performance.

Table 7
The Results of MWA-CNN and Interpretable Space-Based Methods Under Four Kinds of Noise.

Noise	MWA-CNN	Wavelet-Kernel-Net (Laplace-ResNet)	EMD-CNN	FFT-CNN	1DCNN
4 dB	**99.70 ± 0.03**	99.46 ± 0.21	99.58 ± 0.13	99.57 ± 0.15	98.10 ± 0.054
0 dB	**99.65 ± 0.11**	98.89 ± 0.34	99.05 ± 0.23	98.83 ± 0.29	96.53 ± 0.68
−2 dB	**98.99 ± 0.14**	97.04 ± 0.37	97.97 ± 0.52	96.66 ± 0.53	92.49 ± 0.65
−4 dB	**98.71 ± 0.21**	94.73 ± 0.63	95.15 ± 0.84	94.30 ± 0.59	89.29 ± 1.68

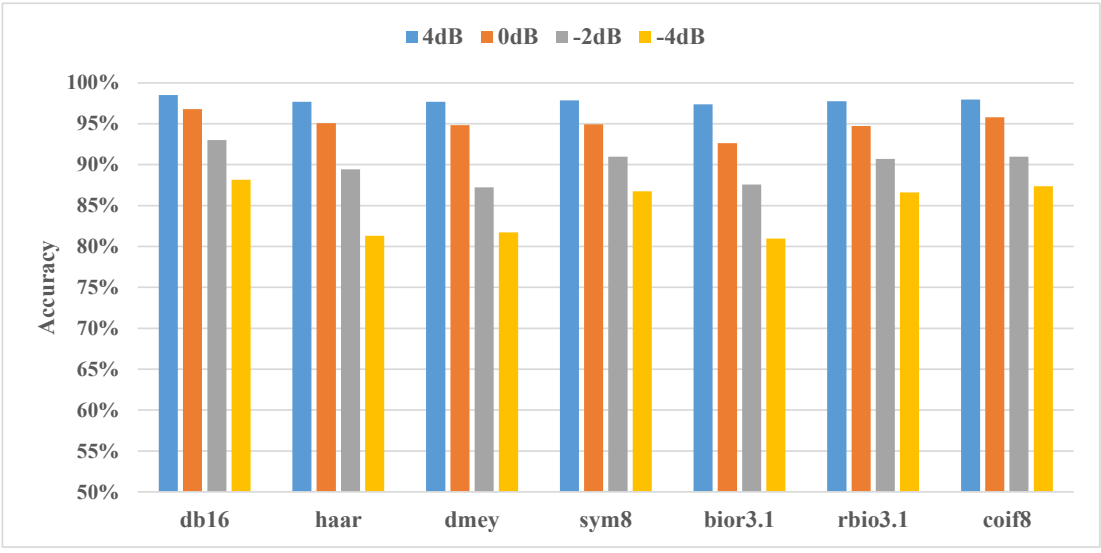


Fig. 16. The Experimental results of MWA-CNN with different wavelet basis functions (SNR = 4 dB, 0 dB, −2dB and − 4 dB).

diagnostic performance of most methods is between 86% and 88%. In general, when the value of N is around 16, the model can obtain relatively good performance.

6. Interpretability analysis

This section explores the feature learning mechanism of the proposed method. Specifically, we first analyze the model's feature

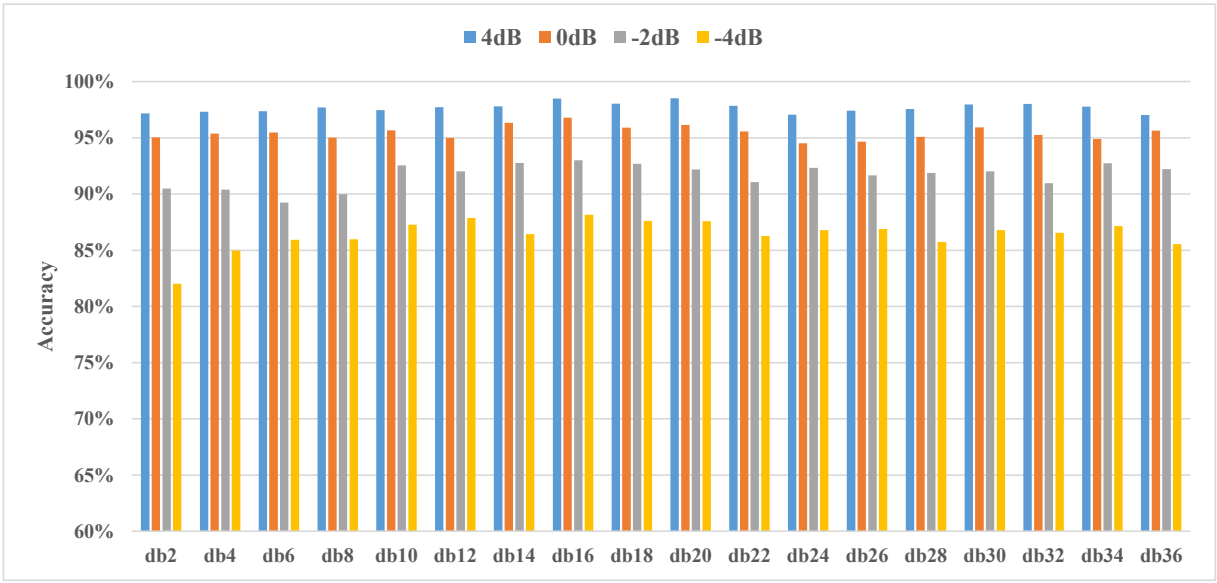


Fig. 17. The Experimental results of MWA-CNN with different db*N* (SNR = 4 dB, 0 dB, −2dB and − 4 dB).

learning preference by visualizing attention weight vectors. Subsequently, the time-domain diagrams, frequency-domain diagrams, and squared envelope spectra of the learned features are shown to explore the learning process of the proposed model.

6.1. *Frequency attention weights visualization*

Fig. 18 shows the MWA-CNN attention weights for two signal samples of category F3 and F4, the experiment uses the HSA bearing dataset and SNR = 4 dB. In MWA-CNN, a total of five FAMs are used, and the weight vectors ($z'_1, z'_2, \ldots, z'_5$) of these FAMs are shown in Fig. 18. Each small grid represents a weight value. The weight value is represented by a color, and the brighter the color, the greater the value. The weight value represents the importance of the corresponding frequency feature. The left side represents the attention weights of low-frequency feature components, and the right side represents the attention weights of high-frequency feature components.

Obviously, the attention module gives different weights to different feature signals, which shows that the attention module is trying to distinguish which information is important and which information is not. According to the bearing fault mechanism, when a bearing has a local fault, the faulty part and other components produce a periodic short-term impact and encourage the bearing system to perform high-frequency free attenuation vibration according to its resonance frequency. In addition, different bearing fault categories will produce different low-frequency fault characteristic frequencies, which are the most critical indicators for distinguishing bearing

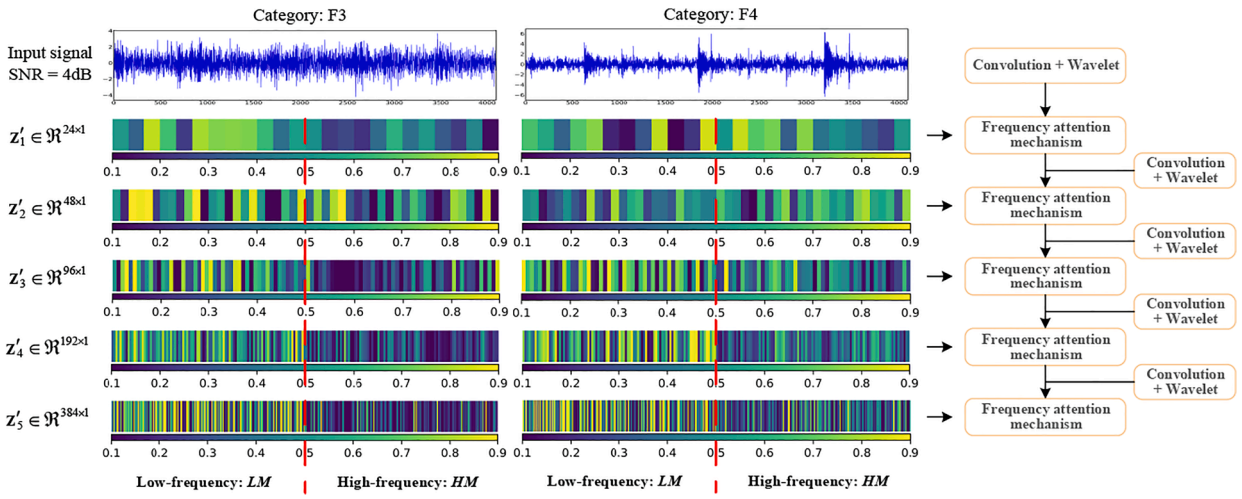


Fig. 18. The attention weights of MWA-CNN for two signal samples of two different classes.

faults. Based on this, we found that CNNs tend to learn fault-related low-frequency features and, when necessary, learn a small number of high-frequency features to achieve accurate fault identification. Generally, the shallow layer of the model is mainly responsible for filtering irrelevant information and retaining important information, and the deep layer of the model is responsible for modeling high-level abstract features. This shows that the network gradually filters and learns the high and low frequencies of the signal in the shallow layers, and then models the fault-related low-frequency features in the deep layers.

6.2. *Visualization of learned features*

To further explore the feature learning mechanism of MWA-CNN, we perform the FFT on the features learned in MWA-CNN and obtain their square envelope spectra. Fig. 19 shows the time-domain and frequency-domain diagrams of the learned features, and their squared envelope spectra. ①, ②, ③, and ④ represent the feature signals obtained from the corresponding positions of the network, respectively. ① indicates the input signal sample.② represents the feature signal output by the first layer convolution module. ③ represents the feature signal output after wavelet transform. ④ is obtained by the inverse wavelet transform of the feature signals output of the first DWA-Layer. The Case Western Reserve University bearing dataset is used to facilitate the visualization of fault characteristic frequencies on the squared envelope spectrum. Due to the ReLU function, the time domain diagram only contains positive values.

Features ② and ④ represent the input and output features of the first DWA-Layer, respectively. As shown in Fig. 19, from the time-domain and frequency-domain diagrams, there is little difference between feature ② and feature ④. The difference between Features ② and ④ is evident in the squared envelope spectrum. The fault characteristic frequency in the square envelope spectrum of feature ④ is clearly visible, but the fault characteristic frequency in the square envelope spectrum of feature ② is not clearly displayed. This shows that through wavelet decomposition and frequency attention learning, important fault-related features are effectively learned and highlighted. This again confirms the usefulness of the proposed method.

To explore the learning process inside DWA-Layer, we have visualized feature ③. Fig. 20 shows the squared envelope spectra of the four feature signals and their corresponding attention weights. These four feature signals are all selected from the low-frequency part. As shown in Fig. 20, the feature signal with a large attention weight generally has a relatively obvious fault characteristic frequency. For example, for feature signals with attention weights of 0.8082, 0.5202, and 0.5965, their fault characteristic frequencies are clearly visible. For the feature signal with a weight of 0.2783, its fault characteristic frequency is not clearly displayed. This shows that DWA-Layer enhances valuable information and suppresses useless information through frequency decomposition and attention mechanisms. In the visualization experiment, the attention mechanism will occasionally give greater weight to the feature signal with an unclear fault characteristic frequency, which may be that the signal contains unknown but essential features. This needs to be further explored in future work.

7. Conclusions

This study aims to deeply integrate the DWT and CNN models to maximize the advantages of these two technologies. To this end, a

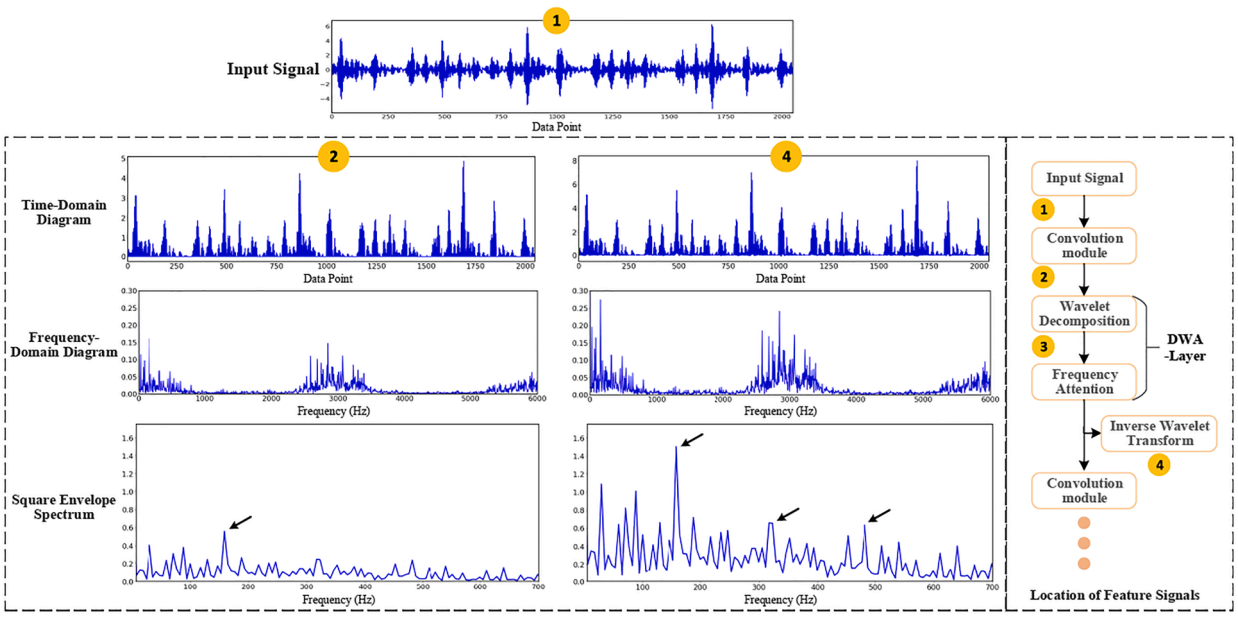


Fig. 19. The time-domain and frequency-domain diagrams of the learned features, and their squared envelope spectra (arrows highlight the fault characteristic frequency).

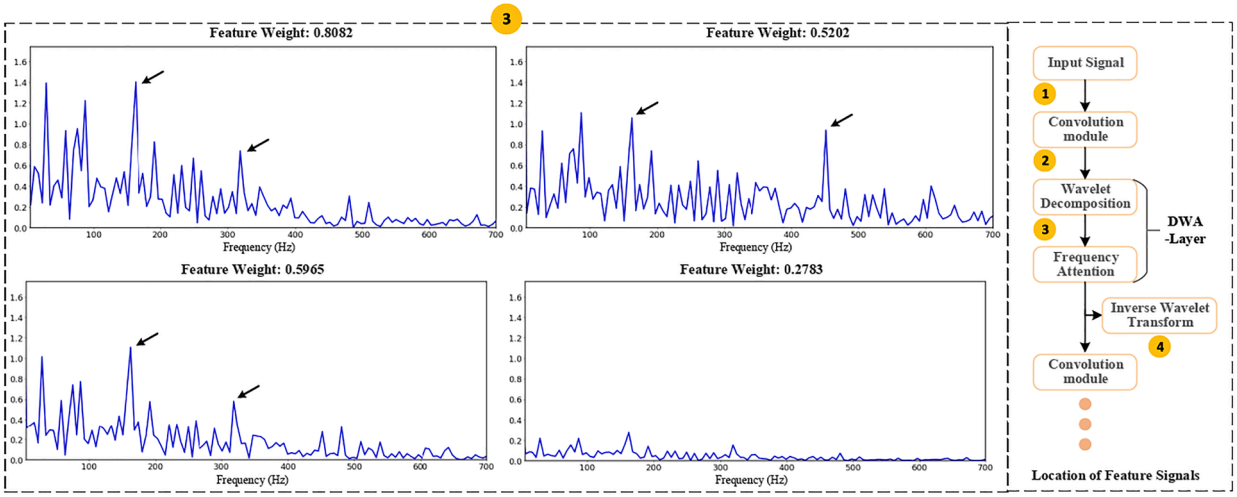


Fig. 20. The squared envelope spectra of the four feature signals and their corresponding attention weights (arrows highlight the fault characteristic frequency).

novel multi-layer wavelet attention CNN for machinery fault diagnosis is proposed. MWA-CNN is mainly composed of multiple DWA-Layers and multiple convolutional layers. DWA-Layer and convolutional layers are stacked alternately. In DWA-Layer, the DWT algorithm is used to decompose the signal into multiple frequency components, and the attention mechanism is used to filter out useful frequency information. The convolutional layer is mainly used to automatically learn useful information from the obtained frequency components. Experimental results show that DWA-Layer can significantly improve the fault diagnosis performance and noise robustness of the CNN model. MWA-CNN has better performance than other deep learning methods. Especially in a strong noise environment, MWA-CNN has excellent performance. Based on DWT, we have analyzed the feature learning mechanism of CNN from the perspective of frequency domain with the help of the attention mechanism, which also provides a direction for the interpretability analysis of the CNN model. In future work, multi-wavelet fusion strategies can be explored and studied to improve the performance of the CNN model further.

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

We have open-sourced our code.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1702400, and in part by Sichuan Province Key Research and Development Program under Grant 23ZDYF0212.

References

[1] M. Cerrada, R. Sánchez, C. Li, F. Pacheco, D. Cabrera, J. Valente de Oliveira, et al., A review on data-driven fault severity assessment in rolling bearings, Mech. Syst. Sig. Process. 99 (2018) 169–196.
[2] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: A review and roadmap, Mech. Syst. Sig. Process. 138 (2020), 106587.
[3] W. Zhang, C. Li, G. Peng, Y. Chen, Z. Zhang, A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load, Mech. Syst. Sig. Process. 100 (2018) 439–453.
[4] D. Peng, H. Wang, Z. Liu, W. Zhang, M.J. Zuo, J. Chen, Multibranch and Multiscale CNN for Fault Diagnosis of Wheelset Bearings Under Strong Noise and Variable Load Condition, IEEE Trans. Ind. Inf. 16 (7) (2020) 4949–4960.
[5] H. Liu, J. Zhou, Y. Zheng, W. Jiang, Y. Zhang, Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders, ISA Trans. 77 (2018) 167–178.
[6] J. Shi, D. Peng, Z. Peng, Z. Zhang, K. Goebel, D. Wu, Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks, Mech. Syst. Sig. Process. 162 (2022), 107996.
[7] T. de Bruin, K. Verbert, R. Babuška, Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks, IEEE Trans. Neural Networks Learn. Syst. 28 (3) (2017) 523–533.
[8] F. Zhou, S. Yang, H. Fujita, D. Chen, C. Wen, Deep learning fault diagnosis method based on global optimization GAN for unbalanced data, Knowl.-Based Syst. 187 (2020), 104837.

[9] H. Wang, Z. Liu, D. Peng, M. Yang, Y. Qin, Feature-Level Attention-Guided Multitask CNN for Fault Diagnosis and Working Conditions Identification of Rolling Bearing, IEEE Trans. Neural Networks Learn. Syst. 33 (9) (2021) 4757–4769.
[10] B. Zhao, X. Zhang, H. Li, Z. Yang, Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions, Knowl.-Based Syst. 199 (2020), 105971.
[11] X. Wang, D. Mao, X. Li, Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network, Measurement 173 (2021), 108518.
[12] T. Jin, C. Yan, C. Chen, Z. Yang, H. Tian, S. Wang, Light neural network with fewer parameters based on CNN for fault diagnosis of rotating machinery, Measurement 181 (2021), 109639.
[13] Z. Liu, H. Wang, J. Liu, Y. Qin, D. Peng, Multitask Learning Based on Lightweight 1DCNN for Fault Diagnosis of Wheelset Bearings, IEEE Trans. Instrum. Meas. 70 (2021) 1–11.
[14] Z. Chen, K. Gryllias, W. Li, Mechanical fault diagnosis using Convolutional Neural Networks and Extreme Learning Machine, Mech. Syst. Sig. Process. 133 (2019), 106272.
[15] H. Han, H. Wang, Z. Liu, J. Wang, Intelligent vibration signal denoising method based on non-local fully convolutional neural network for rolling bearings, ISA Trans. 122 (2022) 13–23.
[16] H. Wang, Z. Liu, D. Peng, Y. Qin, Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis, IEEE Trans. Ind. Inf. 16 (9) (2020) 5735–5745.
[17] Z. Yang, J. Zhang, Z. Zhao, Z. Zhai, X. Chen, Interpreting network knowledge with attention mechanism for bearing fault diagnosis, Appl. Soft Comput. 97 (2020), 106829.
[18] D. Zhou, Q. Yao, H. Wu, S. Ma, H. Zhang, Fault diagnosis of gas turbine based on partly interpretable convolutional neural networks, Energy 200 (2020), 117467.
[19] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, et al., WaveletKernelNet: an interpretable deep neural network for industrial intelligent diagnosis, IEEE Transactions on Systems, Man, and Cybernetics: Systems 52 (4) (2022) 2302–2312.
[20] J. Chen, Z. Li, J. Pan, G. Chen, Y. Zi, J. Yuan, et al., Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review, Mech. Syst. Sig. Process. 70–71 (2016) 1–35.
[21] M.M.M. Islam, J. Kim, Automated Bearing Fault Diagnosis Scheme Using 2D Representation of Wavelet Packet Transform and Deep Convolutional Neural Network, Comput. Ind. 106 (2019) 142–153.
[22] R. Chen, X. Huang, L. Yang, X. Xu, X. Zhang, Y. Zhang, Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform, Comput. Ind. 106 (2019) 48–59.
[23] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, Z. Zhang, Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform, Comput. Ind. 113 (2019), 103132.
[24] Y. Zhang, K. Xing, R. Bai, D. Sun, Z. Meng, An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image, Measurement 157 (2020), 107667.
[25] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.
[26] M. Lin, Q. Chen and S. Yan, "Network In Network," [Online]. Available: https://arxiv.org/abs/1312.4400.
[27] Y. Wu and K. He, "Group Normalization," in Proc. ECCV, 2018, pp. 3-19.
[28] M. Reza Asadi Asad Abad, H. Ahmadi, A. Moosavian, M. Khazaee, M. Ranjbar Kohan and M. Mohammadi, "Discrete wavelet transform and artificial neural network for gearbox fault detection based on acoustic signals," Journal of Vibroengineering, vol. 15, no. 1, pp. 459-463, 2013.
[29] O.N. Oyelade, A.E. Ezugwu, A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram, Sci. Rep. 12 (1) (2022) 5913.
[30] F. Cotter, "Uses of Complex Wavelets in Deep Convolutional Neural Networks," University of Cambridge, 2019.
[31] A.P. Daga, A. Fasana, S. Marchesiello, L. Garibaldi, The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data, Mech. Syst. Sig. Process. 120 (2019) 252–273.
[32] L. Wen, X. Li, L. Gao, Y. Zhang, A new convolutional neural network-based data-driven fault diagnosis method, IEEE Trans. Ind. Electron. 65 (7) (2018) 5990–5998.
[33] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016, pp. 770-778.
[34] Karen Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. International Conference on Learning Representations, 2015.
[35] C. Lessmeier, J.K. Kimotho, D. Zimmer and W. Sextro, "Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification," in Proc. European Conference of the Prognostics and Health Management Society, 2016, pp. 5-8.