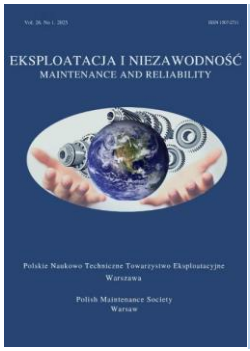


Eksploatacja i Niezawodnosc – Maintenance and Reliability

Volume 27 (2025), Issue 1

journal homepage: http://www.ein.org.pl

Article citation info:
Zhang J, Kong X, Han T, Cheng L, Li X, Liu Z, Research on a Lightweight Multi-Scale Feature Fusion and its Fault Diagnosis Method for Rolling Bearing with Limited Labeled Samples, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2025: 27(1) http://doi.org/10.17531/ein/192235

Research on a Lightweight Multi-Scale Feature Fusion and its Fault Diagnosis Method for Rolling Bearing with Limited Labeled Samples

Indexed by: Web of Science Group

Jiqiang Zhang[a], Xiangwei Kong[a,b,c,*], Taorui Han[a], Liu Cheng[a], Xueyi Li[d], Zhitong Liu[a]

[a] School of Mechanical Engineering and Automation, Northeastern University, China
[b]Key Laboratory of Vibration and Control of Aero-Propulsion System, Ministry of Education, Northeastern University, Shenyang 110819, China
[c]Liaoning Province Key Laboratory of Multidisciplinary Design Optimization of Complex Equipment, Northeastern University, Shenyang 110819, China
[d]College of Mechanical and Electrical Engineering, Northeast Forestry University, China

Highlights

- The proposed method aims to realize fault diagnosis on limited labeled samples.
- A multi-scale depth-separable convolutional neural network is proposed.
- An improved feature soft threshold denoising module is introduced.
- The framework is simpler and clearer, with high robustness and generalization ability.
- The proposed method is more suitable for complex practical engineering scenarios.

Abstract

Convolutional neural networks(CNNs) show significant potential for bearing fault diagnosis. However, traditional CNNs face challenges such as poor noise resistance, high computational complexity, reliance on extensive samples, and limited generalizability. As a result, this paper proposes WDSC-Net, a lightweight, multiscale feature fusion method, focusing on limited labeled fault samples. Initially, a wide kernel convolutional is employed, aiming to reduce parameters and computational complexity. Next, features are fed into a 1×1 convolutional layer reduces feature dimensionality. Subsequently, leveraging the benefits of depth-separable convolution (DSC) allows the separation of spatial and channel features, constructing four convolutional layers of varying scales to amplify the nonlinear fault representation. Finally, an improved feature soft-threshold denoising module is introduced for global feature denoising. Validation on CWRU and MCDS datasets shows that the WDSC-Net method exhibits superior generalizability and noise resistance compared to typical deep-learning fault methods.

Keywords
rolling bearing, limited labeled samples, multi-scale feature fusion, depth-separable convolution network, soft threshold

1. Introduction

Rotating machinery plays a crucial role in industrial production, necessitating timely and accurate fault diagnosis to ensure equipment safety and seamless production. The working environment of rotating equipment inevitably leads to degradation. With the continual increase in the rotational speed, scale, and complexity of equipment, there is a growing demand for higher accuracy and efficiency in fault diagnosis [1]. Following the failure of a rotating component, its vibration amplitude undergoes changes, which are subsequently transmitted to the equipment shell through multiple paths. Consequently, monitoring the vibration signal provides accurate state information about the equipment. However, in real engineering scenarios, the vibration characteristics of rotating machinery are susceptible to various factors, such as high background noise and changes in working conditions. Moreover, equipment in these scenarios often works solely under normal

(*) Corresponding author. J.Zhang (ORCID: 0000-0002-2856-0929) freedomzhangneu@163.com, X. Kong (ORCID: 0000-0001-7042-4368)
E-mail addresses: shawnkongneu@163.com, T. Han (ORCID: 0000-0001-7966-1677) 1425086685@qq.com, L. Cheng (ORCID: 0000-0002-7094-3835) lchengneu@l63.com, X. Li (ORCID: 0000-0003-0335-0594) lixueyiphm@163.com, Z. Liu (ORCID: 0009-0002-9318-5376) lzt9711@163.com

conditions, resulting in an abundance of data on the normal health state but insufficient data on faults. This scarcity of labeled, high-quality data presents a significant challenge in training deep learning models for reliable fault diagnosis. Despite extensive research efforts by domestic and international scholars in recent years, there is an ongoing need to develop and enhance the existing fault diagnosis techniques for rotating machinery.

With the increasing development of fault diagnosis technology, scholars have introduced various traditional intelligent fault diagnosis methods involving manual extraction of original signal features [2-5]. Although these methods effectively diagnose faults, manual extraction of fault features is challenging, particularly when phase information is lost. Additionally, the scarcity of fault samples, coupled with strong background noise and complex working conditions, hinders the rapid extraction of fault features and accurate identification of health states [6]. Hence, there is a compelling need to develop an intelligent algorithm to achieve rapid and accurate fault type identification. In recent years, deep learning has emerged as the primary approach for intelligent fault diagnosis owing to its robust automatic feature extraction capabilities, which are currently popular. Among these approaches, CNNs are representative algorithms in deep learning. In contrast to methods relying on a priori knowledge and signal processing, CNNs can directly handle time signals and adaptively extract the vibration characteristics of rotating equipment layer by layer. This process simplifies fault diagnosis and reveals the intrinsic relationship between the original data and the nonlinear information of each network [7]. Xu et al. [8] proposed a bearing fault diagnosis method based on deep CNNs and random forests. Fan et al. [9] introduced an adaptive deep CNNs fault diagnosis method for rolling bearings. Gong et al. [10] conducted an in-depth analysis of hyperparameter selection and training techniques for CNNs to improve the generality and operability of the model structure. Ye et al. [11] proposed a deep convolutional neural network that fuses features from convolutional layers across different channels and scales through kurtosis and residual-based learning.

The aforementioned research has demonstrated improved diagnostic results in the field of fault recognition. However, structural deficiencies have persisted. CNN-based diagnostic models are evolving toward increased depth, with models such as VGGNet [12], AlexNet [13], and ResNet [14] being designed to enhance performance by adding network layers. An increased number of layers impedes gradient flow, heightens the difficulty of parameter optimization, and increases the susceptibility of the model to overfitting and vanishing gradients. Therefore, it is necessary to develop a lightweight diagnostic network that ensures diagnostic effectiveness while minimizing the number of model parameters. In fault diagnosis, researchers have proposed various lightweight models to tackle diverse challenges. Fang et al. [15] proposed LEFE-Net, employing dynamic and separable convolutions with 2D time-frequency feature mapping for rapid and accurate fault diagnosis. Deng et al. [16] introduced HS-KDNet, addressing fault diagnosis with imbalanced data. Lu et al. [17] proposed a lightweight transfer learning framework for rolling bearing diagnosis based on knowledge extraction. This framework transfers features from a large teacher model to a smaller student model, reducing computational and parameter overhead. Xiong et al. [18] designed a multi-branch deep residual network to enhance the nonlinear characteristics and parameter count of bearing fault diagnosis models. Liu et al. [19] proposed an LSTM cell structure with forgetting gates for low-latency, lightweight recurrent neural networks in machinery fault diagnosis. Cui et al. [20] proposed a lightweight rolling bearing fault diagnosis method using Gramian Angular Field (GAF) and Coordinated Attention (CA), significantly reducing the computational complexity of the model. Beyond fault diagnosis models, Qin et al. [21] explored two-stage detectors for real-time generalized detection and proposed ThunderNet, a lightweight solution for mobile devices. Iandola et al. [22] introduced the SqueezeNet architecture to reduce server communication and bandwidth needs in autonomous driving systems.

Although the aforementioned lightweight models have shown significant results in fault diagnosis, their design process typically requires extensive experimentation and trial error. Optimizing model design for specific diagnostic tasks requires a blend of unique expertise and experience. Furthermore, the portability of these models is generally poor. To better address the challenges of limited resources and real-world application requirements, in-depth research and appropriate model design strategies are essential. Crucially, the research results are

usually validated using datasets with ample samples. In real engineering scenarios, the equipment works under normal conditions for extended periods, making it time-consuming and expensive to collect a large number of high-quality fault data samples. Additionally, existing studies have struggled to effectively mitigate the impact of noise and variable operating conditions on models. Particularly in environments with high noise levels, the identification and classification effectiveness of the model must be improved. Li et al. [23] proposed a signal preprocessing method that transforms 1D vibration signals into 2D grayscale maps. They combined this with an adaptive anti-noise convolutional neural network (AA-CNN) for high-performance fault diagnosis in noisy environments. Dong et al.[24] transformed 1D time series into 2D images using short-time Fourier transform. They employed a parallel large kernel attention mechanism to achieve high accuracy in fault feature extraction across different dimensions on noisy datasets. Li et al. [25] proposed a periodic convolutional neural network that captures noisy vibration signal features under various conditions. They inserted a periodic convolutional module, based on a generalized short-time anti-noise correlation method, before the backbone network. Fan et al. [26] proposed an enhanced anti-noise correlation method for bearing condition monitoring. They validated it by simulating the degradation process and using two rolling bearing accelerated degradation datasets. Overall, despite the strong potential of deep learning methods in fault diagnosis, future research should focus on developing more efficient and robust lightweight models with limited fault samples to better meet real industrial needs.

The problem of fault diagnosis with limited labeled samples can be grouped into three solutions. The first involves training a powerful feature extractor to maintain recognition accuracy regardless of the sample quantity. The second is a data enhancement method that transforms and expands existing training samples. The third method utilizes the parameters of a trained model on a large-scale dataset, adapting to the challenge of limited labeled samples by fine-tuning or adjusting certain parameters in the migration learning method. Chen et al. [27] applied transfer learning to rolling-bearing fault classification, effectively addressing complex data distribution differences under various working conditions. Yang et al. [28] utilized an adaptive auxiliary classifier generative adversarial network (GAN) to expand the sample capacity by generating fault data, demonstrating the effectiveness of the method across multiple datasets. Zhang et al. [29] proposed Meta-GAN, a generalized model for a limited number of labeled samples. Although these methods exhibit improved performance when limited labeled samples are used, they still have some shortcomings. Adversarial generative networks are challenging to train, making them prone to vanishing or exploding gradients. Transfer learning struggles to measure the differences in data distributions between the source and target domains, leading to negative transfer phenomena.

This study effectively solves the problem of fault discrimination for rotating machinery with the limited labeled samples by introducing a lightweight multiscale feature fusion WDSC-Net method. The method consists of three core modules: a wide kernel convolution module, a multiscale feature fusion module based on depth-separated convolution, and a semisoft thresholding feature denoising module.

First, we collected the raw acceleration signals of the bearing by utilizing acceleration sensors and input the raw vibration signals under mixed working conditions into the network. The raw data are passed through the wide-kernel convolution module, which assists the model in better understanding the features of the input data and capturing the bearing fault characteristics more effectively. Compared to using multiple small convolutional kernels, the wide-kernel convolution module can process more input information in a single convolutional layer, thus reducing the number of parameters and depth of the model. Next, the extracted features are downscaled by 1×1 convolutional layers to further reduce the computational complexity while performing deeper feature extraction. Subsequently, taking advantage of the depth-separable convolution, four convolutional layers with different scales (LMSFMs) are constructed to enhance the nonlinear representation of limited fault samples and to facilitate more efficient convergence. In addition, a soft-thresholding module is embedded at the end of the LMSFM for hierarchical feature denoising, avoiding the complexity of manually setting thresholds. Finally, the bearing faults are diagnosed using a softmax classifier through the global maximum pooling layer. Through experimental validation in simulating a real engineering noise environment and various types of bearing

faults under different working conditions, the proposed method has a significant advantage over classic deep learning methods in diagnosing bearing faults with limited labeled samples.

The main contributions of the paper are summarized as follows:

(1) This study aims to address the problem of realizing high-precision fault diagnosis in scenarios with limited labeled samples. Unlike CNN fault diagnosis methods, which focus on a large number of high-quality fault samples, our work is closer to real industrial scenarios.

(2) To reduce model parameters and deployment costs in real industrial scenarios, an efficient, lightweight, and multiscale fault diagnosis framework was proposed. The framework has a simple and clear hierarchical structure, strong robustness, and excellent generalization ability.

(3) To evaluate the effectiveness of the proposed method, a series of tasks was devised on two bearing datasets, CWRU and MCDS. The experimental results show that the proposed method outperforms the current popular fault diagnosis methods.

The subsequent sections of the paper are organized as follows: Section 2 presents the research objectives and scope of limited labeled sample learning. Section 3: describes the proposed WDSC-Net in detail. Section 4 provides a comprehensive overview of the experimental dataset. Section 5: demonstrates the superiority of the method through comparisons with typical deep learning methods on both public and self-built datasets. Finally, conclusions and future research directions are presented in Section 6.

2. Study purpose and scope of limited labeled samples

Intelligent diagnosis of deep neural networks relies on ample fault monitoring data [30]. Increased training data adequacy and a diverse range of fault types in the training set are directly correlated with higher accuracy in the intelligent diagnosis model. Nevertheless, practical engineering scenarios pose challenges in establishing an ideal dataset for training diagnostic models for the following four reasons.

(1) Limited fault data in real-world engineering. In real-world engineering scenarios, the equipment typically works under normal conditions, resulting in a scarcity of fault data. Although a condition monitoring system with multiple sensors can collect equipment status data, the majority of these data are normal health data. This scarcity of fault data hinders the effective training and testing of intelligent diagnostic models.

(2) Challenges in replicating real-world faults in laboratory settings. Replicating fault data in a laboratory setup identical to real engineering scenarios is challenging and expensive. Typically, obtaining fault data in a laboratory involves purchasing a simulation laboratory bench and inducing various faults in-house. However, some faults, such as those caused by gear gluing, are difficult to simulate manually.

(3) Disconnection between computer simulations and real working conditions. Data from computer simulation software are detached from the actual working conditions, making it challenging to simulate the impact of failure data on the working environment and conditions.

(4) Challenges in equipment intelligence and data collection in real engineering. Actual engineering scenarios face challenges owing to the low level of equipment intelligence, the absence of state data collection equipment, and the lack of awareness among field operators regarding monitoring data collection.

In brief, fault diagnosis in real engineering scenarios inherently involves limited labeled samples, as indicated in the literature [31-32], for which sample numbers range from tens to hundreds. This paper explores the feature extraction of the WDSC-Net method, emphasizing rotating machinery fault diagnosis. The dataset includes 20, 40, 60, 80, 100, 200, and 400 samples for each class of signals.

3. Theory of the WDSC-Net method

In recent years, integrating the advantages of diverse networks, exploring potential feature learning mechanisms in diagnostic models, and enhancing the learning ability of fault features have posed significant challenges. With the increasing size, speed, integration, and automation of mechanical equipment, the pursuit of high-precision diagnosis has led to increasingly complex network structures and growing model parameters, directly impacting the scalability of the model and diagnostic costs. Therefore, the authors developed a lightweight fault diagnosis framework that aims to meet the requirements of computational efficiency and practical application. Additionally, in real-world engineering scenarios, there is often a need for efficient diagnostic models, and lightweight fault diagnosis

models can maintain high performance while reducing computational resources and memory consumption. More importantly, lightweight networks are more suitable for practical deployment. In industrial scenarios, deployment costs and real-time considerations are crucial factors, and the simplification and efficiency of lightweight networks make them more suitable for embedded systems, edge devices, or real-time applications. The lightweight fault diagnosis network we constructed based on limited samples exhibits strong robustness and generalizability compared to traditional CNN structures.

The network structure proposed in this study accommodates one-dimensional raw vibration signals as inputs without requiring operations, such as a time-frequency domain transformation. For one-dimensional vibration signals, the width of the first convolutional kernel layer significantly influences fault diagnosis performance [33]. WDSC-Net consists of a wide convolutional layer, a 1×1 convolutional layer, two LMSFM modules based on a depthwise separable convolutional network (DSC), and a soft-threshold feature denoising module.

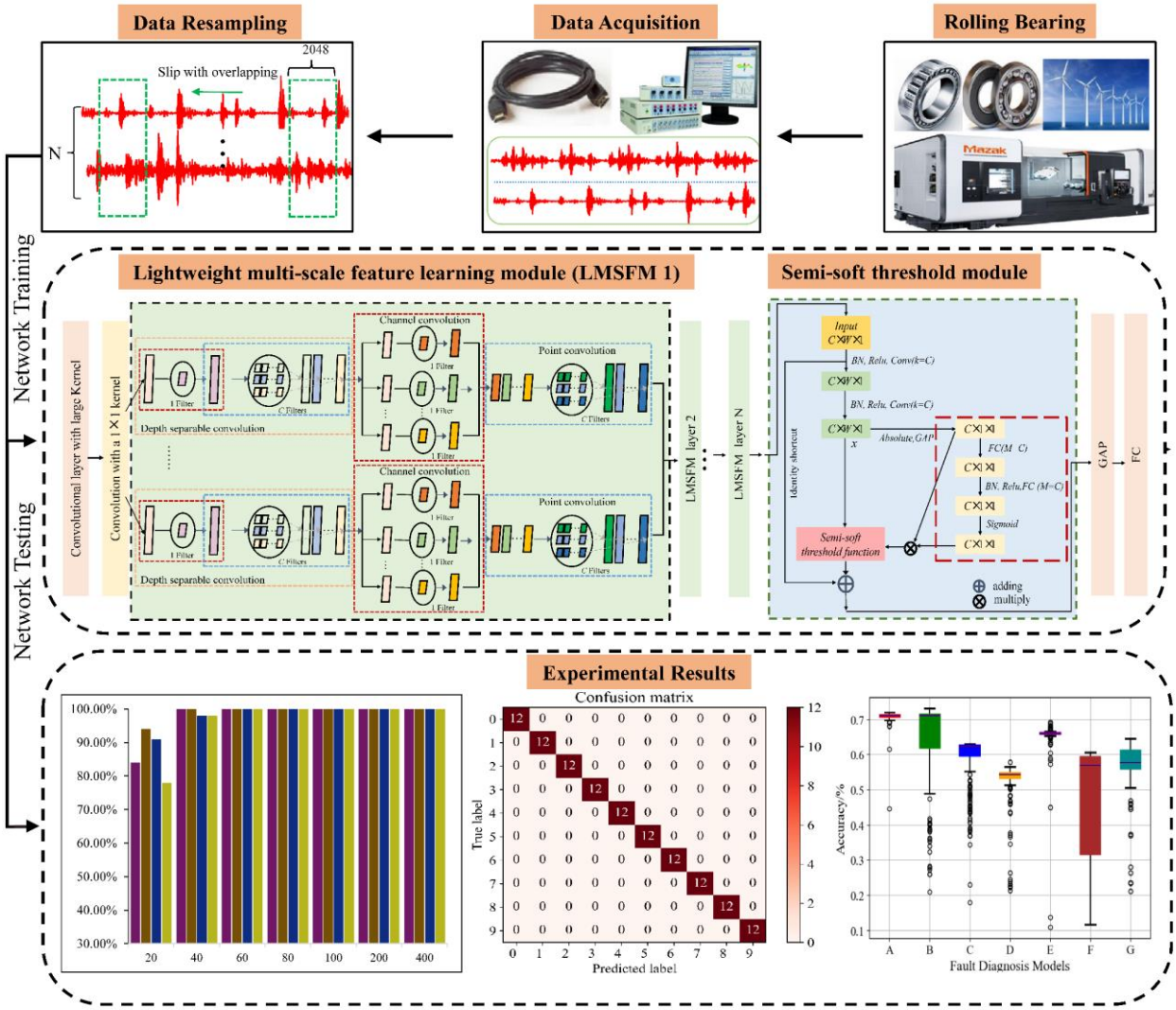


Fig. 1. Fault diagnosis framework of the WDSC-Net method.

The wide convolutional layer has a relatively large kernel size, enabling it to capture a broader range of fault features in the input signal. This helps model the deep information of vibration signals, enhances the perception of global information, and reduces the interference of noise on fault features. The 1× 1 convolutional layer is used for feature dimensionality reduction, reducing the computational complexity and number of parameters of the network[34]. It helps retain important features and eliminate redundant information, making the model is more lightweight while maintaining sensitivity to key

information and improving network efficiency. The LMSFM module consists of a DSC with four kernels of different sizes. The DSC network decomposes the filters of standard convolutional layers into channel convolution and point convolution layers [35]. Pruning smaller weight parameters during the point convolution process reduces the number of parameters, enhancing the computational efficiency and parameter utilization of the model. The LMSFM is a multiscale feature fusion network designed to enable the simultaneous processing of feature information at different scales. By combining fault information at different scales, the nonlinear expression capability and comprehensive capture of input data features are enhanced, thereby improving the diagnostic performance of the model. Importantly, rotating mechanical systems may contain various fault modes that manifest at different scales. Multiscale feature fusion allows the network to better adapt to and capture diverse fault modes, increasing the applicability and generalizability of the model. The soft-threshold feature denoising module is used to remove noise from the signal. This module helps improve the robustness of limited labeled samples, ensuring that the network focuses more on real fault features, thereby enhancing diagnostic accuracy. The fault diagnosis framework of WDSC-Net is illustrated in Fig. 1.

The convolutional layer extracts features from the original data through convolutional operations. This operation involves multiplying the input data by a convolution kernel, adding bias constants, and using a sliding window with a specific step size. The set of learnable kernel functions is a crucial parameter in the feature extraction stage, directly influencing the quality of the output features [36]. In this study, rather than decreasing the model parameters by employing a small convolutional kernel, we constructed a multiscale feature fusion network based on deeply separable convolutions, which have fewer model parameters, for faulty bearing identification and classification.

The raw vibration signals contain time-dependent information. Because faults are localized, the impact response to vibration may not occur at every time point, leading to varying response times for different faults. Hence, a relatively small convolution kernel may not capture the complete range of information associated with the fault impact. In this study, we initially employ a wide kernel to broaden the sensory field of view and extract short-term features. This approach aims to uncover more global information across different states and to identify segments affected by faults. The representation is as follows.

$$Z^{l+1}(i,j) = \mathrm{Conv}[Z^l \otimes \mathbf{W}^{l+1}](i,j) + b = \sum_{K_l=1}^{K} \sum_{x=1}^{f} \sum_{y=1}^{f} \left[Z_K^l(s_0 i + x, s_0 j + y) \mathbf{W}_k^{l+1}(x,y) \right] + b \quad (1)$$

where Z^l and Z^{l+1} represent the input and output feature maps of the $l+1$ layer, respectively. $\boldsymbol{W}$ is the weight matrix of the convolution kernel, i, j are the coordinate points of the feature maps, b is a bias constant, K is the number of channels in a convolution layer, f is the width of the convolution kernel, and s_0 is the size of the step.

Each convolutional layer in the network uses the ReLU as an activation function, offering several advantages: (1) Nonlinear strength. When the input x is greater than zero, the activation function outputs x; otherwise, it is zero. Stacking multiple hidden layers with a strongly nonlinear activation function enables the model to achieve more complex nonlinear mappings. This helps the model adapt to a wider range of data distributions and complex relationships, ultimately improving the feature extraction capability of the model. (2) Mitigation of the vanishing gradient problem. Compared to activation functions such as sigmoid and tanh, ReLU helps alleviate the vanishing gradient problem. During backpropagation in the model, gradients pass through each layer, and sigmoid and tanh tend to have gradients close to zero for large or small inputs, leading to ineffective gradient propagation. ReLU, with a constant gradient for positive inputs, avoids the vanishing gradient problem, facilitating more effective gradient updates and learning. (3) Sparse activation nature. The ReLU has a sparse activation property, outputting zero when the input is negative. This property helps the network focus more on learning fault features, enhancing the ability to extract important information from limited labeled samples. (4) High computational efficiency. Its mathematical expression is $f(x) = max(0, x)$. Compared to the sigmoid and tanh activation functions, the ReLU activation function has simpler calculations, and its straightforward computational structure results in a lower computational burden during both forward and backward propagation, contributing to faster training. The ReLU function is represented by the following equation:

$$\sigma ReLU = max(0, x) \quad (2)$$

The output values are then normalized by batch normalization.

$$y = BN(\sigma_{relu}\{Conv(\mathbf{W}, x)\}) \quad (3)$$

where y is the output feature map after normalization. Whereas traditional convolutional operations involve differentiation computations, DSC conducts a convolution operation for each channel independently [37]. In other words, each channel was convolved individually using a single convolutional kernel. Subsequently, all the channels are stacked, and the number of output feature maps is adjusted to match the number of channels in the input layer. This deep convolution process enhances the differential computation of feature information across spatial locations.

$$Z^{l+1}(i,j) = \mathrm{DWConv}[y \otimes \mathbf{W}^{l+1}](i,j) + b = \sum_{F}^{f} \mathbf{W} f * y_{(i,j)} + f \quad (4)$$

$$z = BN(\sigma_{relu}\{Z^{l+1}(i,j)\}) \quad (5)$$

where z is the output feature map of the deep convolution.

Simultaneously, to effectively utilize information from different feature maps at the same spatial location, this study employed point-by-point convolution (with a convolution kernel size of 1×1) to weigh and combine the feature maps in the depth direction. This approach facilitates the exchange and integration of cross-channel information.

$$\hat{Z^{l+1}}(i,j) = \mathrm{PWConv}[z \otimes \mathbf{W}^{l+1}](i,j) + b = \sum_{F}^{f} \mathbf{W} f * z_{(i,j)} \quad (6)$$

$$z = BN\left(\sigma_{relu}\{\hat{Z^{l+1}}(i,j)\}\right) \quad (7)$$

Assuming that the size of the input feature map is H, the number of input channels is S, the size of the convolution kernel is D_f, and the number of convolution kernels is K, Then, the following parameters are computed for conventional convolution and depth-separable convolution:

$$H \times S \times K \times D_f \quad (8)$$

$$H \times S \times D_f + H \times S \times K \quad (9)$$

Thus, the parameter ratio of the depth separable convolution to the conventional convolution can be expressed as:

$$\frac{H \times S \times D_f + H \times S \times K}{H \times S \times K \times D_f} = \frac{1}{K} + \frac{1}{D_f} < 1 \quad (10)$$

According to Eq. (10), the 1D depth-separable convolution significantly decreases the number of computations and improves the convergence efficiency of the model. The depth separable convolution has a significant advantage over the traditional 1D convolution in terms of the number of computational parameters.

The structure of the DSC layer is shown in Fig. 2.

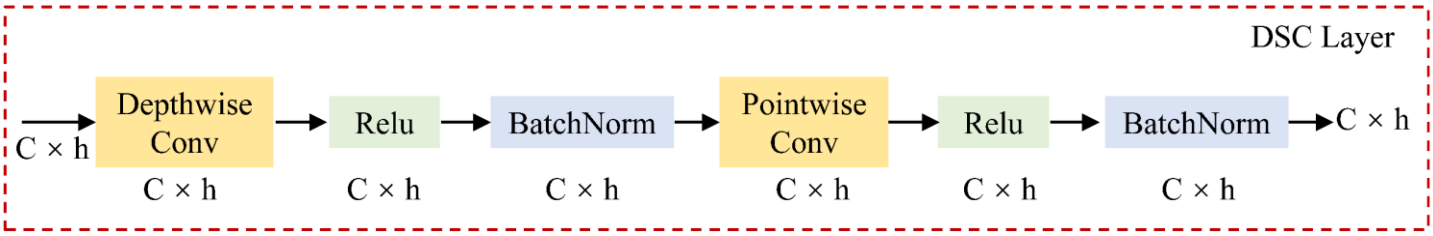


Fig. 2. The structure of the DSC layer.

Multiscale feature fusion has been widely adopted in recent years [38-39]. Given the outstanding recognition performance of the DSC, this study explored feature fusion across different scales. This approach aims to enrich fused features with more comprehensive fault information, facilitating fault recognition and classification tasks under the constraint of limited labeled samples. Moreover, the bearing vibration signals studied here exhibit nonlinearity and nonsmoothness. In addition, owing to the impact of the acquisition environment and working conditions, even bearings in the same healthy state may exhibit a certain data distribution offset. Therefore, employing a multiscale feature fusion method allows analysis of the multiscale features of the original signal.

The key to building a multiscale feature fusion model based on DSC lies in designing the convolutional kernel size. Convolutional kernels of different sizes can be crafted to learn effective features at various scales. A small convolutional kernel is suitable for learning features with closer intervals, whereas a large kernel is effective for learning features with more distant correlation intervals. In this study, convolutional kernels of sizes (1×3), (1×5), (1×7), and (1×9) were designed. However, the use of larger convolutional kernels increases the computational effort of the overall network structure. Therefore, in this study, we initially conducted a 1×1 convolution with the following advantages. (1) Feature dimensionality reduction. This operation reduces the feature dimension and decreases the

number of computational parameters. (2) Enhanced feature fusion information. Utilizing convolution kernels of different sizes enriches feature fusion information. (3) Improved model expression. Deepening the network enhances the capacity of the model for expression. Considering these factors, the design of the multiscale feature extraction module based on DSC is depicted in Fig. 3.

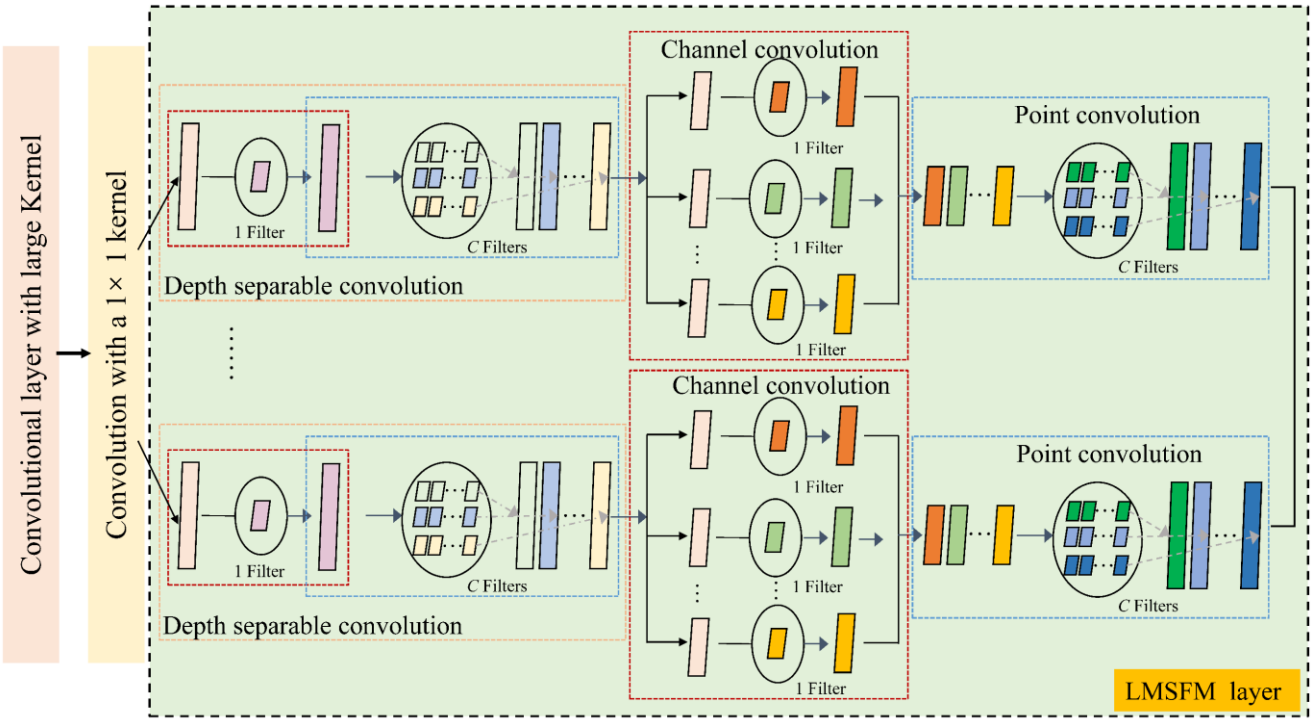


Fig. 3. Multiscale feature fusion module based on the DSC.

In practical engineering scenarios, the acquisition of vibration signals often results in contamination by varying degrees of noise, leading to interference from fault information in the signals and consequently affecting the accuracy of the diagnostic results. Therefore, denoising the original signal is a focal point in fault diagnosis research. Over the past 20 years, soft thresholding has typically been employed as a key step in signal denoising [40]. This technique generally transforms the useful information of the original signal to positive or negative features while converting noise information close to zero. However, determining useful and noisy information often requires significant signal-processing expertise and setting a reasonable threshold can be challenging. Moreover, the optimal threshold depends on the working conditions.

To address the aforementioned problem, this study integrates the soft threshold method with a deep learning approach and incorporates the soft threshold as a nonlinear transformation layer into the unit [41]. The design principle of the semisoft threshold module is based on the demand for enhancing and denoising fault signals. This module performs soft threshold processing on specific global features to reduce noise components in the signal and emphasize the significant features of the fault signal. Additionally, this module exhibits better smoothness than other threshold processing methods, aiding in retaining subtle yet crucial information in the signal. The semisoft threshold module includes key parameters such as global feature input, global average pooling, scale factor calculation, and soft threshold processing. This approach introduces a new perspective for problem solving and circumvents the laborious and arbitrary task of manually setting a threshold value. The soft threshold function can be expressed as

$$y = \begin{cases} sgn(x) \cdot (|x| - \tau), |x| \geqslant \tau \\ 0, \qquad\qquad\quad |x| < \tau \end{cases} \tag{11}$$

where x denotes the input. y represents the output. τ is the threshold. As evident from Eq. (11), after soft threshold function processing, the data maintain good continuity and denoising in the signal while preserving the effective signal. The learning process of the soft-threshold module is illustrated in Fig. 4.

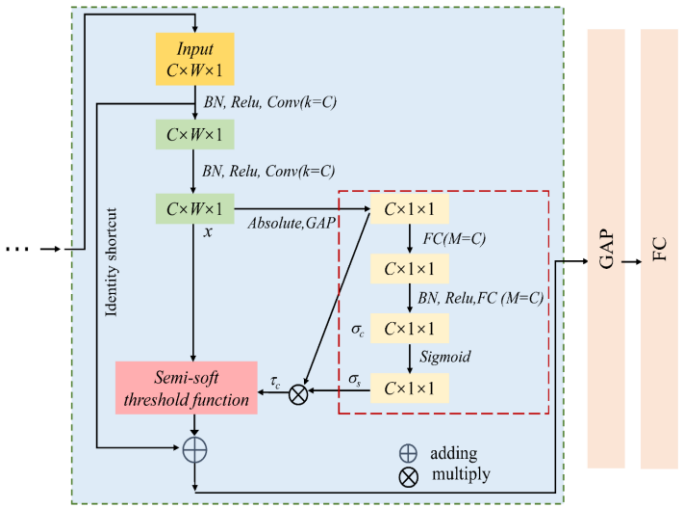


Fig. 4. The learning process of the soft threshold module.

First, global average pooling (GAP) was applied to the absolute values of the features to obtain a one-dimensional vector. Next, a one-dimensional vector is passed through a two-layer fully connected (FC) network to obtain the scale parameter σ_c. In the second FC layer, the number of neurons was equal to the number of channels. Finally, a sigmoid function was applied at the end of the second FC layer to scale the parameter to (0,1). The scale parameter σs can be expressed as

$$\sigma s = \frac{1}{1+e^{-z_c}} \tag{12}$$

where z_c and σs represent the features of the neuron and the scale parameter of the c layer, respectively.

Assuming that the width and height of the feature map are denoted as i and j, respectively, and the number of channels is c, the soft threshold can be defined as

$$\tau_c = \sigma s \cdot \underset{i,j,c}{average}\left|X_{i,j,c}\right| \tag{13}$$

The pooling layer processes the regional features obtained from the soft-threshold module as statistical features. The aim of this operation is to reduce the number of parameters. The average pooling layer is expressed as

$$p_j(i) = \underset{(i-1)W+1\le t\le iW}{average} y_j(t) \tag{14}$$

Here, $y_j(t)$ represents the value of the t -th neuron in the j -th frame, $t \in [(i-1)W+1, iW]$. W is the width of the current region, and $p_j(i)$ represents the value of the neuron in the pooling layer.

Our analysis of multiple experiments revealed that the softmax classification function in full connectivity outperformed the other functions. This is expressed as follows.

$$Softmax(x) = exp\big(p_j(i)\big) / \sum_{\bar{M}}^{i=1} exp\big(p_j(i)\big) \tag{15}$$

where M represents the number of categories. In accordance with the WDSC-Net model, cross entropy is employed as the loss function, which is defined as follows.

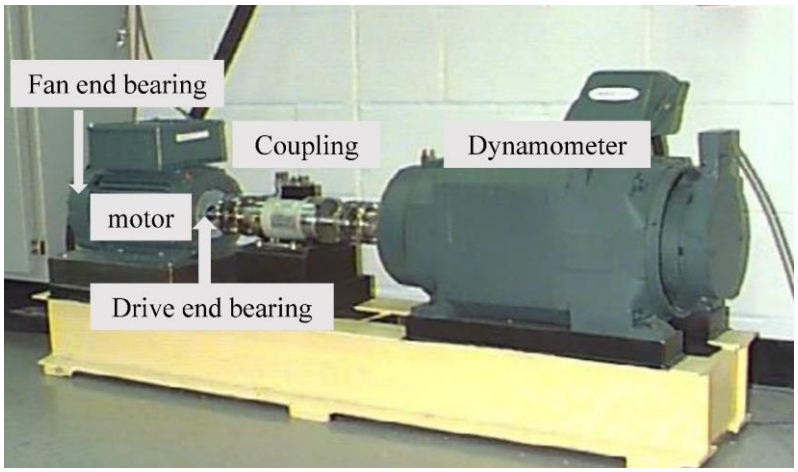
$$Z = Softmax\big(p_j(i)\big) \tag{16}$$

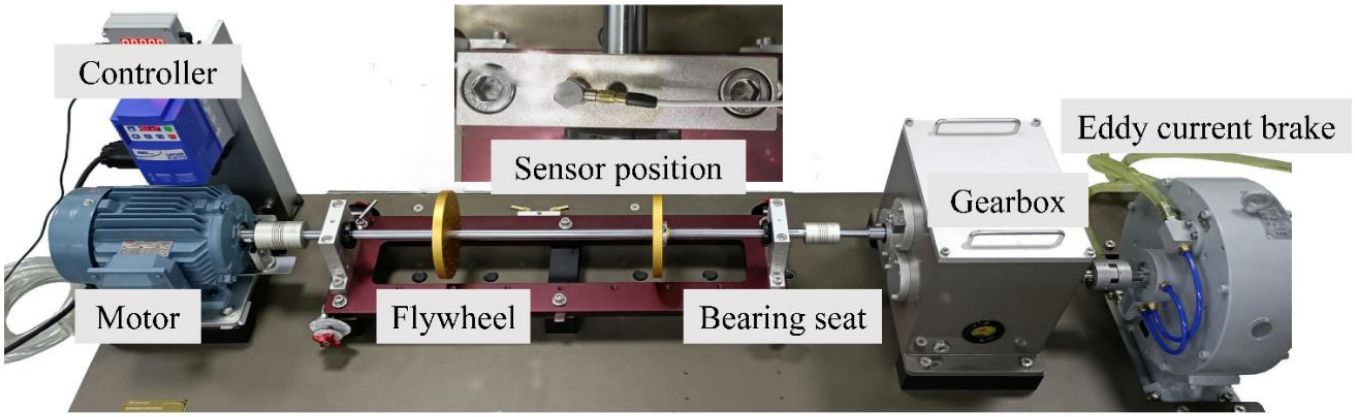
$$L = -\frac{1}{T}\sum_{\bar{T}}^{i=1} y_i \, ln \, Z_i \tag{17}$$

where T is the number of samples in the training set, Z_i denotes the classification result of the ith sample, and y_i is the actual label of the ith sample.

4. Construction of the experimental dataset and parameter sensitivity analysis

4.1 Construction of the experimental data





(b)

Fig. 5. Experimental setup. (a) depicts the CWRU bearing test rig, while (b) illustrates the MCDS bearing test rig.

This paper validates the validity and generalizability of WDSC-Net using the Case Western Reserve University (CWRU) bearing dataset [42] and bearing data obtained from the MCDS integrated failure simulation testbed. Fig. 5 illustrates the test setup for both datasets.

Table 1. Description of the CWRU subdataset.

Datasets				Fault types	Fault diameter (mm)	Labels
F1	F2	F3	F4			
0 hp	1 hp	2 hp	3 hp	Normal	0	0
0 hp	1 hp	2 hp	3 hp		0.18	1
0 hp	1 hp	2 hp	3 hp	Ball	0.36	2
0 hp	1 hp	2 hp	3 hp		0.54	3
0 hp	1 hp	2 hp	3 hp		0.18	4
0 hp	1 hp	2 hp	3 hp	Inner Race	0.36	5
0 hp	1 hp	2 hp	3 hp		0.54	6
0 hp	1 hp	2 hp	3 hp		0.18	7
0 hp	1 hp	2 hp	3 hp	Outer Race	0.36	8
0 hp	1 hp	2 hp	3 hp		0.54	9

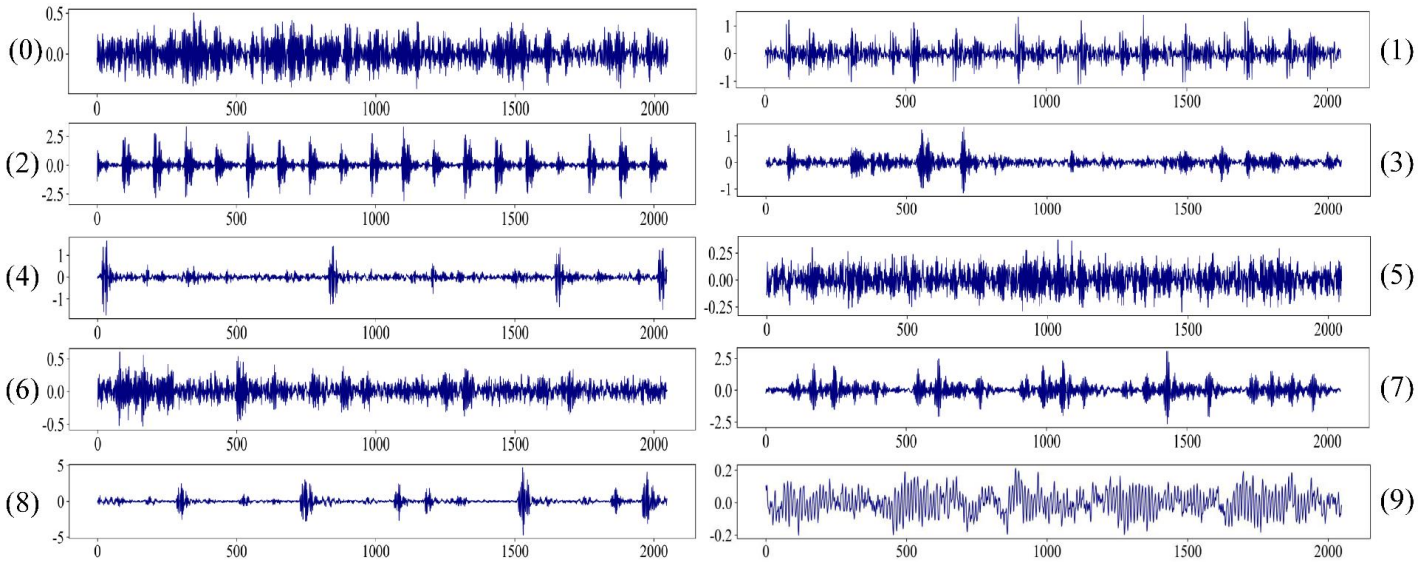


Fig. 6. Time domain diagram of the CWRU signal.

The CWRU experimental setup comprises four components: a motor, a torque encoder, a dynamometer, and control

electronics. These components facilitate the collection of vibration signals from drive and fan-side bearings. The focus of this study was a drive-side bearing (SKF-6205) with a sampling frequency of 12 kHz. The experimental setup induced single-point faults in the bearings using EDM, where the fault locations included the inner ring, outer ring, and rolling faults, each with fault bores of 0.18, 0.36, and 0.54 mm, respectively. The dataset was generated by collecting data at four different loads (0-3 hp), resulting in nine fault states and a normal state for each load. The corresponding labels are 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The performance of the model was evaluated under various working conditions by creating four subdatasets, as detailed in Table 1. Specifically, datasets F1, F2, F3, and F4 correspond to 0 hp, 1 hp, 2 hp, and 3 hp, respectively. Time-domain diagrams of the raw vibration signals are depicted in Fig. 6.

The MCDS comprehensive fault diagnosis simulation test bench is composed of seven parts: a controller, motor, bearing seat, turntable, rigid rotor shaft, gearbox, and eddy current brake. The bearing type is ER-12K, with a sampling frequency of 10240 Hz, a sampling time of 30 s, and fault positions, including the inner ring, outer ring, and rolling faults. Vibration signals were collected under four different working conditions at speeds of 900, 1200, 1500, and 1800 r/min. This process results in three types of fault states and a normal state for each speed; the corresponding labels are 0, 1, 2 and 3, as presented in Table 2. Datasets F5, F6, F7, and F8 corresponded to the data at 900, 1200, 1500, and 1800 r/min, respectively. Time-domain diagrams of the original vibration signals are shown in Fig. 7.

Table 2. Description of the MCDS subdataset.

Datasets				Fault types	Fault diameter (mm)	Labels
F5	F6	F7	F8			
900 r/min	1200 r/min	1500 r/min	1800 r/min	Normal	0	0
900 r/min	1200 r/min	1500 r/min	1800 r/min	Ball	0.15	1
900 r/min	1200 r/min	1500 r/min	1800 r/min	Outer Race	0.30	2
900 r/min	1200 r/min	1500 r/min	1800 r/min	Inner Race	0.20	3

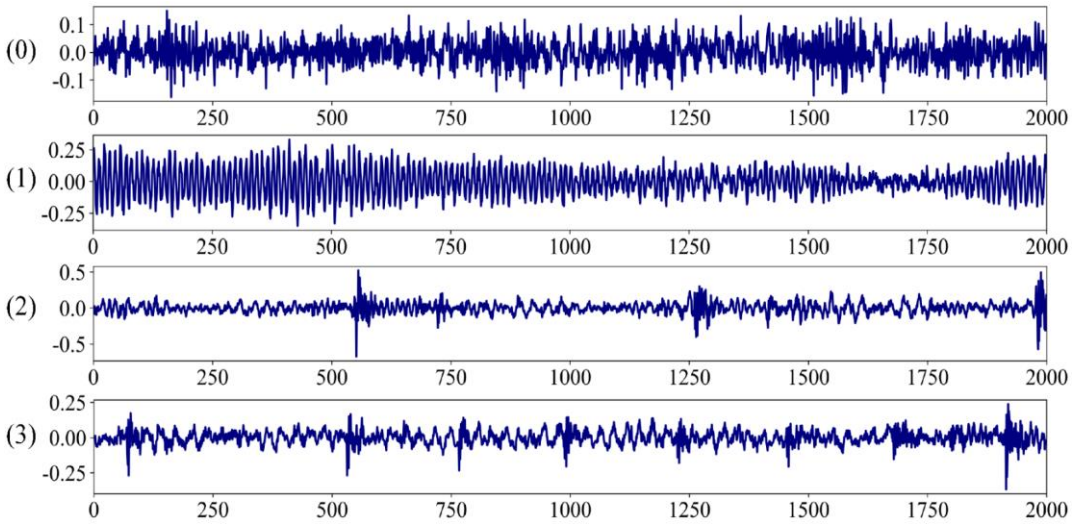


Fig. 7. Time domain diagram of the MCDS signal.

To ensure completeness of the fault information in each data sample, we set the length of each sample to 2048 points. Continuous data points are selected from the original data not involved in training; 80% of the data points are randomly assigned to the training set, while 20% are assigned to the test set. In addition, all the datasets were normalized. The Adam optimizer was used to adjust the learning rate based on the historical gradients. This optimization technique helps the model minimize the loss function rapidly and accurately, enhancing both training effectiveness and model generalizability.

4.2 Parameter sensitivity analysis.

We selected key parameters, including the learning rate, batch size, wide kernel convolution size, number of multiscale feature modules, and number of epochs, for the model sensitivity analysis. The detailed analysis is as follows:

Table 3. Network parameter selection of WDSC-Net.

	Learning rate	Batch size	Wide kernel convolution size	Multiscale feature modules	Epochs	Average accuracy (%)	Time (s)
Choice	0.001	32	200	2	100	100	62
	0.01					99.70	62
	0.1					51.04	62
	0.001	16	200	2	100	97.30	123
		8				95.40	239
		4				95.10	495
	0.001	32	100	2	100	98.10	62
			50			97.30	62
			25			93.30	62
	0.001	32	200	1	100	92.60	34
				3		99.38	91
				4		90.63	116
	0.001	32	200	2	50	84.30	32
					30	82.02	26
					20	81.95	14

Learning rate sensitivity analysis: We selected learning rates of 0.001, 0.01, and 0.1 for our experiments. The results show that the convergence time is similar across these learning rates. However, model performance decreases significantly with higher learning rates, particularly from 0.01 to 0.1. Batch size sensitivity analysis: We selected batch sizes of 32, 16, 8, and 4 for our experiments. The results indicate that larger batch sizes achieve higher accuracy in a shorter time, while smaller batch sizes increase training time and slightly decrease accuracy. This suggests that larger batch sizes are more effective in practical applications. Convolutional kernel size sensitivity analysis: We selected kernel sizes of 200, 100, 50, and 25 for our experiments. The results show that larger kernel sizes significantly improve model accuracy. Notably, larger kernel sizes enhance the feature extraction capability. Sensitivity analysis of the number of multiscale feature modules: We selected 1, 2, 3, and 4 multiscale feature modules for the experiments. The results show that the model performs best with 2 modules. More than 2 modules lead to overfitting, indicating that increasing the number of modules requires caution to avoid performance degradation due to excessive complexity. Epoch sensitivity analysis: Increasing the number of epochs improves the model accuracy but also significantly increases the training time. Thus, balancing accuracy and training time is necessary in practical applications.

Based on the sensitivity analysis of key parameters such as learning rate, batch size, kernel size, number of multiscale feature modules, and epochs, we selected a learning rate of 0.001, a batch size of 32, a kernel size of 200, 2 multiscale feature modules, and 100 epochs. This combination optimizes model performance and ensures effective training.

5. Analysis and discussion of the results

Various experiments were conducted to validate the diagnostic performance of the proposed models. First, experiments are performed on the CWRU and MCDS datasets to compare the accuracy with limited labeled samples. Second, in real engineering scenarios, vibration signals are often contaminated by varying degrees of noise, leading to obscuring of the fault information in the signal. To enhance the anti-interference ability of the model and prevent data overfitting, we introduced Gaussian noise with a signal-to-noise ratio of -6~6 dB to the limited labeled samples in the CWRU and MCDS datasets. Third, considering the complex and variable working conditions of mechanical equipment, significant differences in signal characteristics arise. Therefore, 12 scenarios were designed in the CWRU and MCDS datasets to assess the generalization performance of the proposed method in identifying bearing damage levels under variable working conditions with a limited labeled sample background. The software environment for all the experiments was the PyTorch framework in Pycharm 2020.2.1, and the hardware environment consisted of an Intel(R)

Core(TM) i7-10700 CPU @ 2.90 GHz processor and an NVIDIA GeForce GTX1660 Ti graphics card.

where the signal-to-noise ratio (SNR) is defined as:

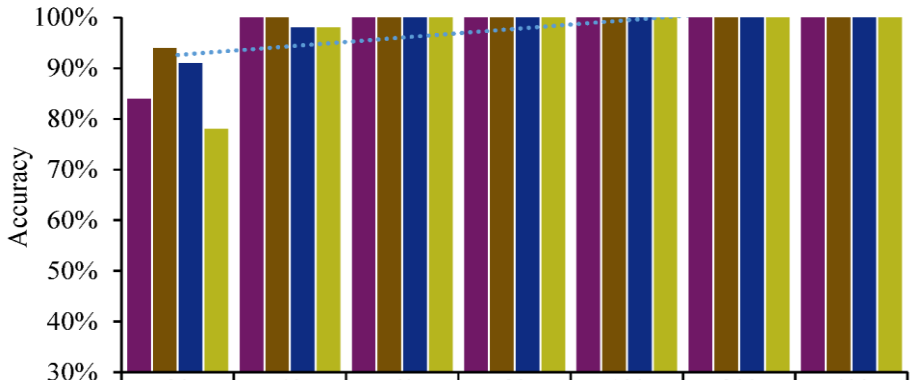
$$\mathrm{SNR}_{\mathrm{dB}} = 10\,log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{18}$$

where P_{signal} and P_{noise} are the signal and noise powers, respectively.

5.1 Impact analysis of model recognition performance with limited labeled samples

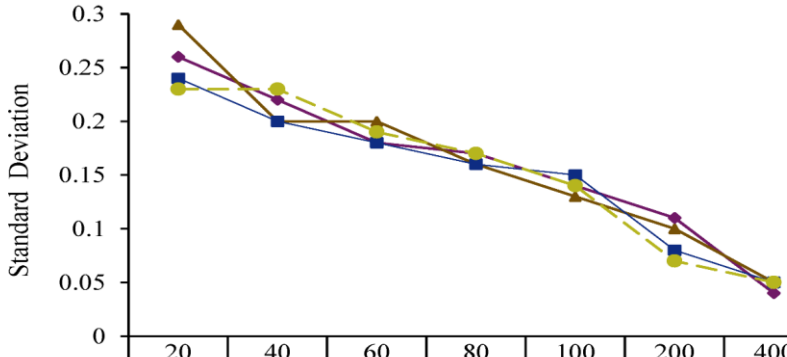
5.1.1 Diagnostic performance in CWRU and MCDS-bearing datasets

In practical engineering scenarios, engineering equipment is typically allowed to work only under normal conditions, resulting in abundant normal and limited fault data. Therefore, this study focused on the CWRU and MCDS datasets to investigate the influence of the number of training samples with fault information on the fault recognition effectiveness of the proposed method. For each class (normal and fault samples), 20, 40, 60, 80, 100, 200, and 400 samples are selected from the four subdatasets. The ratio of the training set to the test set was 8:2. In this paper, the purpose of recording the standard deviation is to measure the variability of the model across 8 datasets. A larger standard deviation indicates poorer stability of the model, while a smaller standard deviation suggests greater stability. The experimental results are shown in Fig. 8.



	20	40	60	80	100	200	400
Dataset F1	84.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Dataset F2	94.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Dataset F3	91.00%	98.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Dataset F4	78.00%	98.00%	100.00%	100.00%	100.00%	100.00%	100.00%

(a)



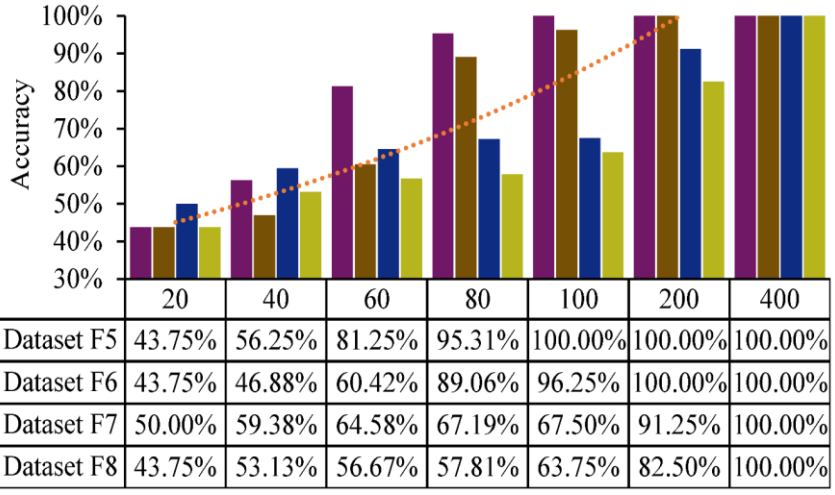
	20	40	60	80	100	200	400
Dataset F1	0.26	0.22	0.18	0.17	0.14	0.11	0.04
Dataset F2	0.29	0.20	0.20	0.16	0.13	0.10	0.05
Dataset F3	0.24	0.20	0.18	0.16	0.15	0.08	0.05
Dataset F4	0.23	0.23	0.19	0.17	0.14	0.07	0.05

(b)

Fig. 8. Effect of limited labeled samples on the recognition accuracy (a) and standard deviation (b) for the CWRU subdataset.

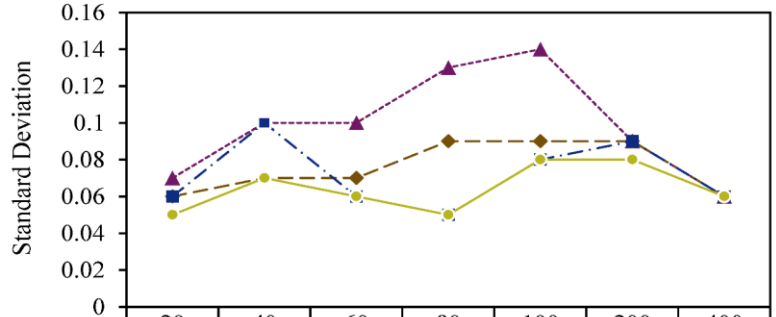
On the CWRU subdataset, the highest accuracy for dataset F2 was 94% when the sample size was 20. However, the

standard deviation is the highest among the four datasets, indicating a degree of fluctuation in the recognition accuracy. Dataset F4 has the lowest recognition accuracy of 78%. As the sample size increases, the recognition accuracy gradually improves and stabilizes. When the sample size is 40, the recognition accuracy has reaches more than 98% on the four subdatasets, and the standard deviation further decreases. When the sample size is greater than 60, the recognition accuracy has reaches 100% on all four subdatasets. Morever, with increasing in sample size, the standard deviation gradually decreases and eventually stabilizes near approximately 0.05.



	20	40	60	80	100	200	400
Dataset F5	43.75%	56.25%	81.25%	95.31%	100.00%	100.00%	100.00%
Dataset F6	43.75%	46.88%	60.42%	89.06%	96.25%	100.00%	100.00%
Dataset F7	50.00%	59.38%	64.58%	67.19%	67.50%	91.25%	100.00%
Dataset F8	43.75%	53.13%	56.67%	57.81%	63.75%	82.50%	100.00%

(a)



	20	40	60	80	100	200	400
Dataset F5	0.07	0.1	0.1	0.13	0.14	0.09	0.06
Dataset F6	0.06	0.07	0.07	0.09	0.09	0.09	0.06
Dataset F7	0.06	0.1	0.06	0.05	0.08	0.09	0.06
Dataset F8	0.05	0.07	0.06	0.05	0.08	0.08	0.06

(b)

Fig. 9. Effect of limited labeled samples on the recognition accuracy (a) and standard deviation (b) for the MCDS subdataset.

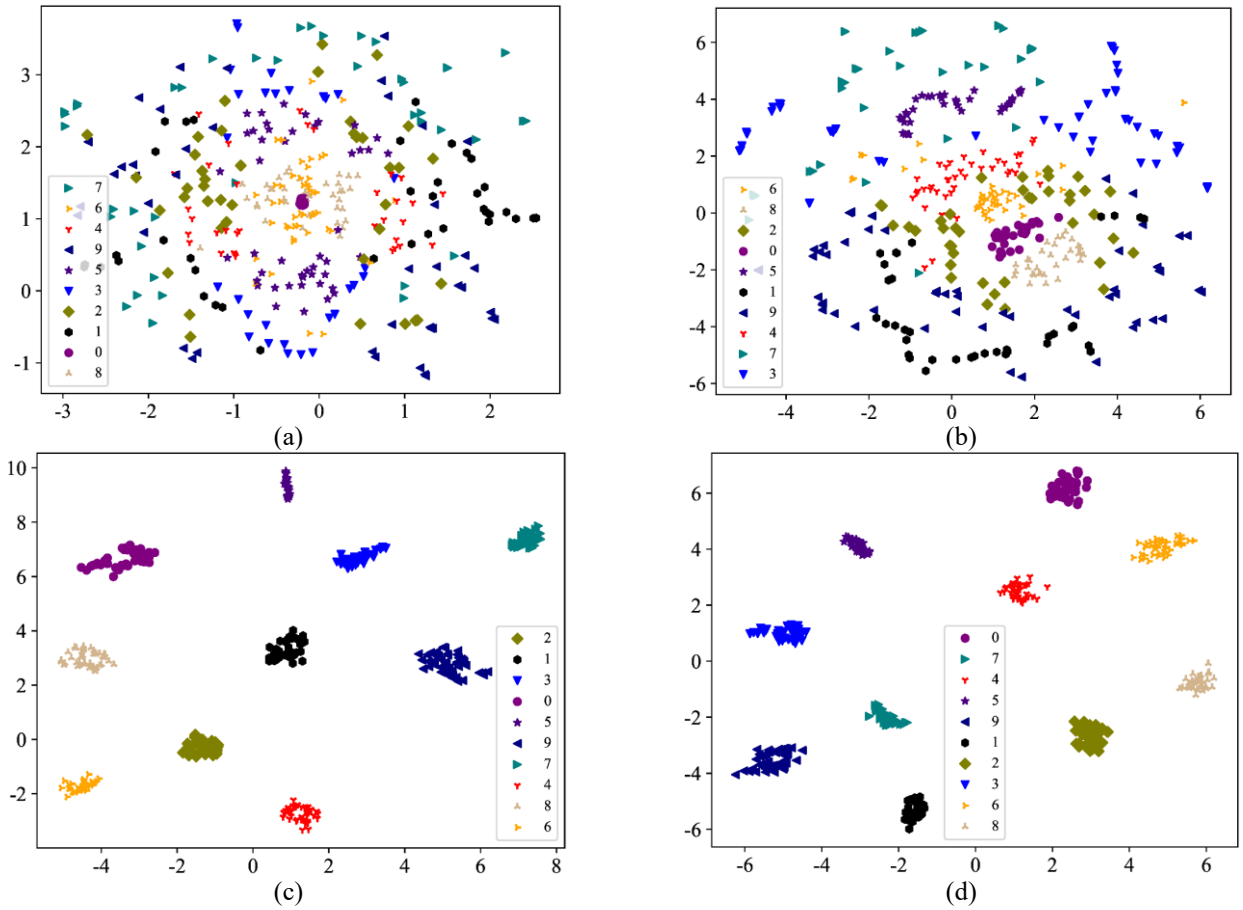
Overall, the overall diagnostic performance on the MCDS subdataset is lower than that on the CWRU subdataset. The recognition accuracy on sub-dataset F7 is 50% when the number of samples is 20. As the number of samples increases, the recognition accuracy improves. When the number of samples was increased to 40, the recognition accuracy on all four subdatasets improved by approximately 10%. When the number of samples increased to 60, the recognition accuracy improved by approximately 20% on F5 and F6, and 5% on F7 and F8. When the number of samples is increased to 80, the recognition performance on the F5 and F6 is better and has reaches approximately 90%, but the diagnostic performance on F7 and F8 is still unsatisfactory. When the number of samples is increased to 200, the recognition accuracy on all four subdatasets has attaines more than 80%. The standard deviations of the four datasets generally increase and then decrease, which indicates that the number of high-quality labeled samples is crucial to the stability of the model.

The above experimental results show that the proposed method has certain applicability and limitations when dealing with limited sample data. When each class is larger than 100 samples, the method can effectively perform fault

discrimination. However, when facing smaller sample data, especially in MCDS when the number of samples per class is between 20 and 60, the method may be affected by data sparsity, which may have led to overfitting of the model to the noisy samples in the training data, leading to the possible challenges of the proposed method in limited sample scenarios. In this case, the proposed method may require more samples for training to adequately capture the feature differences between each class. We plan to integrate multiple classifiers or employ techniques such as transfer learning to reduce the data distribution differences between various domains, improve the performance and generalization ability of the model to better cope with the challenge of minimum small-sample data, and provide more reliable fault diagnosis solutions for practical engineering applications.

To further enhance the interpretability of the model, we randomly selected datasets F2 and F5 and performed t-SNE visualization of the wide kernel convolutional layer, the LMSFM1 and LMSFM2 modules based on depth-separable convolution, and the fully connected layer (FC). The results are as follows:

As observed in Fig.10 (a) and (e), the fault data are more dispersed after passing through the wide-kernel convolutional feature extraction layer, showing no obvious clustering or clear demarcation between classes. This indicates that the wide-kernel convolutional layer has limited feature extraction capability. Fig.10 (b) and (f) show that when the fault data pass through the LMSFM1 module, the data distribution improves. Some classes start to form more obvious clusters, and the dividing line between labels becomes clearer. This indicates that the LMSFM1 module enhances feature extraction. Fig.10 (c) and (g) reveal that after the LMSFM2 module, data point clustering becomes more apparent. Different classes form clearer clusters, with more concentrated data points and more distinct demarcation lines. This indicates that the LMSFM2 module performs better in feature extraction and can better separate data points of different classes. Fig.10 (d) and (h) demonstrate that after the fully connected layer, data point clustering is very obvious. Different types of faults form clear clusters, with distinct demarcation lines between various labels. This indicates that the fully connected layer performs well in feature extraction.



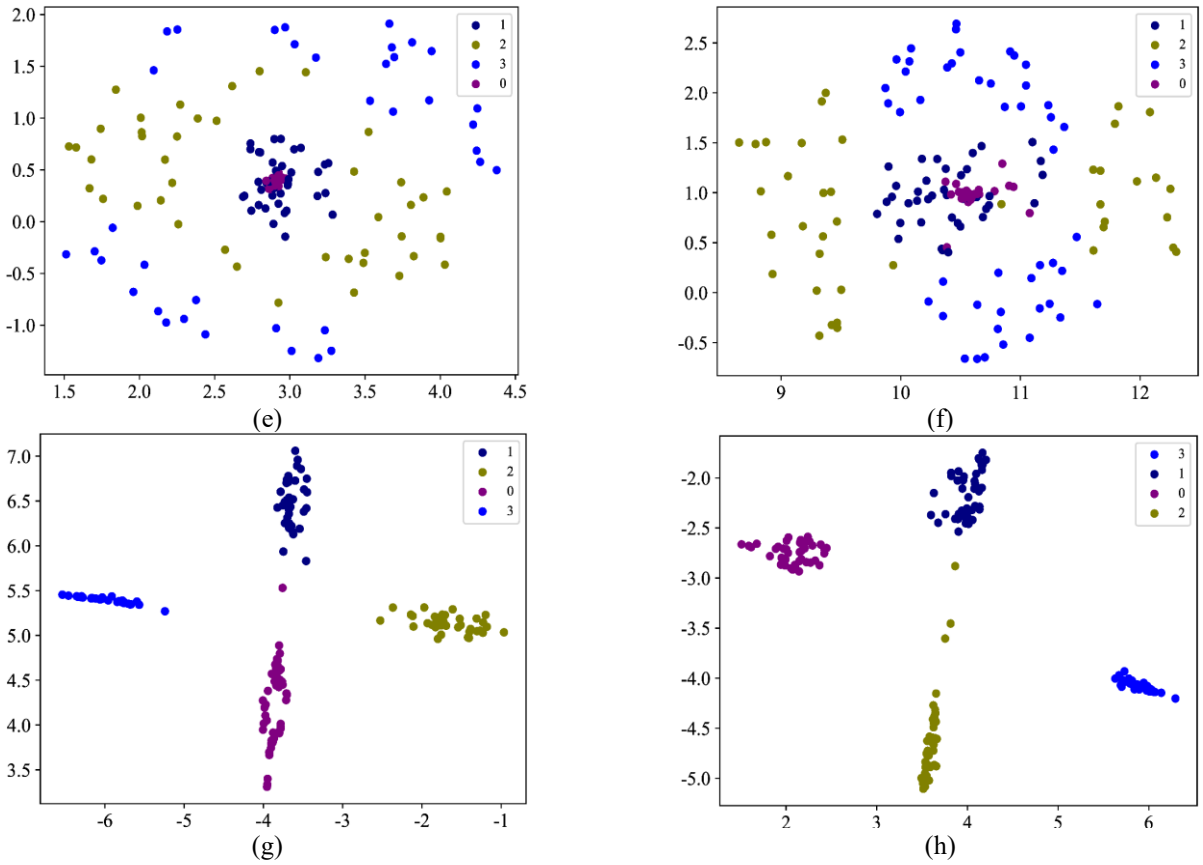


Fig.10. Visualization of the intermediate feature layer t-SNE for the WDSC-Net method.

5.1.2 The underlying reasons for performance gaps in CWRU and MCDS-bearing datasets

The experiment results on the CWRU and MCDS datasets revealed significant differences in diagnostic performance. To determine the underlying reasons, we performed a comparative analysis of the features of these two datasets. We believe that the difference in dataset features may be one of the reasons for the difference in diagnostic performance.

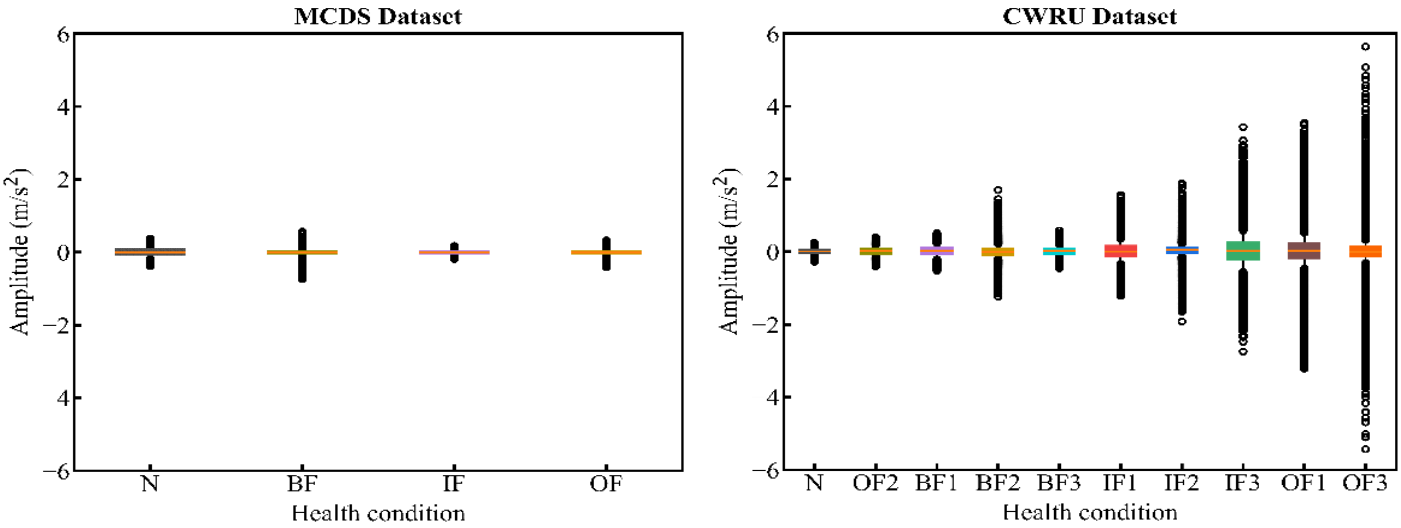


Fig. 11. Box diagrams of the dataset MCDS (F1)and dataset CWRU(F5).

First, the relative cleanliness and the lower noise level of the CWRU dataset are acknowledged in the literature [43]. To illustrate the relative simplicity of the task in the CWRU dataset in greater depth, we chose to compare the characteristics of the data distributions in dataset F1 and dataset F5, and plotted the boxplots for 10,000 adjacent data points. In each health state,

the top and bottom of the boxplots represent the upper and lower quartiles of the data, respectively, i.e., containing 50% of the data. As shown in Fig. 11, the absolute values of the data amplitudes in the MCDS-bearing dataset are all less than 1 m/s², and the fluctuations in the data between different classes are small. On the other hand, in the CWRU dataset, the absolute value of the data amplitude fluctuates greatly with the change in classes, and the highest value is more than 5 m/s². In addition, the absolute value of the data amplitude varies greatly within the same fault class. This suggests that in the CWRU dataset, it is easier to discriminate between different classes of samples, and a smaller number of limited labeled samples may be necessary to attain better diagnostic performance, whereas, in the MCDS dataset, the challenge is greater.

5.2 Validation of the effectiveness of the proposed method against typical deep learning methods

5.2.1 Fault identification accuracy of different models

To verify the performance of the proposed method under limited labeled samples, the samples for the CWRU and MCDS subdatasets were set to 60 and 400 in each class, respectively. This involves training the model with 480 and 1280 samples in each subdataset and using 120 and 320 samples, respectively, to validate the performance of the proposed method. The proposed model was compared and analyzed with nine classical deep-learning fault diagnosis models. Model A represents the proposed WDSC-Net. Model B is the ACNN method [44], an interference-resistant model that introduces the attention mechanism into the feature extraction layer and exhibits excellent generalizability. Model C is the WDCNN [45], which is a wide convolutional kernel deep convolutional neural network that demonstrates good variable load adaptation in strongly noisy environments. Model D is a deep convolutional neural network consisting of two sets of ResNet modules in tandem, aiming to address the issues of gradient vanishing and model degradation through the introduction of a residual module. Model E is the CNN-LSTM, which is a network comprising a convolutional layer and a long short-term memory network designed to capture long-term dependencies in fault data. Model F is the CNN, which is a lightweight neural network consisting of a convolutional layer and a pooling layer. Model G is an improved Let-5 [46] that reduces the distributional variance of inputs and mitigates overfitting by adding a BN layer after each convolutional layer. Model H is a GAN-CNN [47] designed to leverage adversarial mechanisms, allowing the generator to continuously improve the realism of the generated data. Simultaneously, the discriminator continually enhances its ability to discern between real and generated data, aiming to achieve the generation of high-quality auxiliary data. Model I is a GAMCAE [48], a novel method that combines multiscale autoencoders (AEs) and generative adversarial networks to extract depth-sensitive features from signals. These features are then integrated with a classifier for fault diagnosis. Model J adopts the hybrid convolutional autoencoder (HCAE) [49], a semi-supervised fault diagnosis method. HCAE, characterized as a hybrid autoencoder, utilizes a softmax classifier to perform direct health condition diagnosis by leveraging encoded features generated through the autoencoder. Model K is a supervised learning algorithm known as the support vector machine (SVM). Its core principle involves segregating different classes of data by identifying an optimal hyperplane that maximizes the margin between data points of different classes. Model L is an enhanced algorithm called the kernel extreme learning machine (KELM), which builds upon the extreme learning machine framework and incorporates kernel functions. It comprises an input layer, a hidden layer, and an output layer. Model M is a decision tree model (DT) that represents decision rules and classification outcomes in a tree-like data structure. Model N is an ensemble learning algorithm known as random forest (RF), which generates final predictions by simultaneously training the dataset with multiple decision trees. It leverages voting mechanisms or averaging to make predictions. The hyperparameter settings are the same for all the above models, and the test results are presented in Table 4.

For both the CWRU and MCDS subdatasets, the proposed method in this study accurately identifies the fault types of rolling bearings, achieving an accuracy improvement of 21% to 30% compared to the average performance of 13 typical fault diagnosis methods.

Table 4. Fault identification accuracy of different models.

Methods	CWRU subdatasets acuracy (%) & std							
	F1		F2		F3		F4	
Proposed	100.00%	0.12	100.00%	0.10	100.00%	0.10	100.00%	0.09
Model B [44]	93.22%	0.29	78.13%	0.28	96.88%	0.20	93.75%	0.18
Model C [45]	96.87%	0.21	90.63%	0.20	98.45%	0.20	96.85%	0.18
Model D	81.20%	0.16	84.38%	0.15	85.41%	0.11	85.42%	0.12
Model E	93.75%	0.27	90.63%	0.28	72.92%	0.17	79.17%	0.22
Model F	89.58%	0.13	87.50%	0.13	87.50%	0.11	93.75%	0.13
Model G [46]	96.35%	0.15	83.85%	0.15	75.00%	0.09	81.25%	0.13
Model H [47]	56.48%	0.13	61.64%	0.16	58.19%	0.11	49.57%	0.14
Model I [48]	92.48%	0.19	96.31%	0.17	94.56%	0.23	90.13%	0.16
Model J [49]	90.45%	0.13	87.10%	0.16	92.89%	0.16	96.12%	0.11
Model K	39.17%	0.10	42.86%	0.07	47.32%	0.10	40.18%	0.07
Model L	18.33%	0.02	16.67%	0.01	16.67%	0.02	12.50%	0.01
Model M	23.32%	0.01	27.50%	0.02	31.77%	0.01	20.00%	0.01
Model N	44.17%	0.05	51.77%	0.03	65.00%	0.06	50.00%	0.05
Ave acc/std	72.53%	0.14	71.36%	0.14	73.04%	0.12	70.62%	0.11
Methods	MCDS subdatasets acuracy (%) & std							
	F5		F6		F7		F8	
Proposed	100.00%	0.07	100.00%	0.11	100.00%	0.06	100.00%	0.05
Model B [44]	98.40%	0.14	89.10%	0.16	98.44%	0.17	98.13%	0.14
Model C [45]	98.44%	0.23	100%	0.23	98.44%	0.24	96.88%	0.24
Model D	98.75%	0.07	96.88%	0.12	98.13%	0.02	93.13%	0.11
Model E	96.25%	0.32	89.34%	0.28	74.38%	0.20	80.63%	0.26
Model F	95.00%	0.08	95.63%	0.14	98.75%	0.09	91.25%	0.07
Model G [46]	95.63%	0.12	93.75%	0.12	91.88%	0.12	86.88%	0.09
Model H [47]	65.63%	0.13	71.88%	0.15	75.00%	0.11	70.31%	0.16
Model I [48]	94.20%	0.10	95.73%	0.15	93.62%	0.17	89.50%	0.13
Model J [49]	88.70%	0.13	92.37%	0.13	95.14%	0.09	90.27%	0.10
Model K	31.88%	0.02	34.38%	0.03	30.94%	0.01	31.56%	0.03
Model L	31.88%	0.04	30.31%	0.02	28.44%	0.02	30.63%	0.01
Model M	39.38%	0.01	38.75%	0.01	33.69%	0.01	28.13%	0.01
Model N	81.88%	0.06	58.43%	0.02	67.38%	0.02	71.56%	0.04
Ave acc/std	79.72%	0.11	77.61%	0.12	77.45%	0.10	75.63%	0.10

Specifically, in the CWRU subdataset, the test accuracy of the proposed method surpasses the average accuracy of all the methods by 27%~30%. In addition to the proposed method, in dataset F1, Model C achieved the highest accuracy of 96.87%, contributing to the highest average accuracy of 72.53% among all the methods. For dataset F2, Model I achieved the highest test accuracy of 96.31%. For datasets F3 and F4, Model C attained the highest accuracies of 98.45% and 96.85%, respectively. Concerning the MCDS subdataset, the accuracy of the proposed method was 21%~25% greater than the average accuracy of all the other methods. Among the four datasets, excluding the proposed method, Models D and C attained the highest test accuracy on datasets F5 and F6, respectively. Models F and B achieved the highest accuracy for datasets F7 and F8, respectively. Model C achieved 100% accuracy in dataset F6. Although the results show a better effect of the proposed method on limited samples, mdoel (K-N) still has unique advantages and applicability, such as stronger stability, and we will endeavor to integrate our method in future studies, to better promote the development of the field of limited-sample fault diagnosis.

5.2.2 The reasons for the low diagnostic performance of the comparison method

The accuracy of fault diagnosis methods using CWRU datasets has generally approached 100% in recent years, but the diagnostic performance of the comparison methods employed in this paper is slightly lower, including the following reasons:

1) A high-quality and sufficient dataset helps the model to learn the data features better, improving the diagnostic performance. The comparison method applied in this paper may differ from that used in previous studies. This study focuses more on the generalization ability of the model on a dataset with limited labeled fault samples and the diagnostic performance in a strong noise background rather than overfitting the training set, which may have been overlooked in previous studies. 2) The adjustment of training strategies may also lead to differences in performance. In this paper, although we refer to the relevant literature for key parameters such as data preprocessing, learning rate, and batch size, the experimental settings in the paper, including feature extraction and model training, are not detailed in the literature. All these factors have some impact on the final diagnosis, especially on the limited sample dataset. In addition, to reduce randomness and obtain a more stable performance evaluation, we conducted several repeated experiments for each method and obtained the mean value as the final result. 3) More importantly, the design of the model architecture is crucial for task adaptability. The architecture adopted in this paper has been carefully designed to be more suitable for dealing with tasks with strong noise and variable working conditions for diagnosis under a limited fault dataset, which is one of the reasons why we obtained better performance on this dataset. Combining all these factors, the method proposed in this paper outperforms the comparative methods on the limited fault sample dataset, whereas the diagnostic performance of the comparative methods may be lower than that of previous studies.

5.2.3 Detailed discussion on the limitations of existing methods and the specific advantages of WDSC-Net

We have provided a more detailed discussion of related work, focusing on the limitations of existing methods and the specific advantages of WDSC-Net. In this paper, we compare the limitations of various machine learning methods, including SVM, DT, RF, and KELM, and deep learning methods, such as Let-5, CNN, CNN-LSTM, ACNN, WDCNN, GAN-CNN, HCAE, and GAMCAE. We highlight the performance advantages of WDSC-Net, especially in scenarios with limited labeled samples. The detailed analysis is as follows.

(1) Discussion of the limitations of machine learning methods

SVM: Requires tuning of key parameters, such as penalty and kernel parameters, often using cross-validation. It is sensitive to noise and struggles with nonlinear problems. Additionally, it has higher time complexity and memory consumption as the data volume and feature number increase. DT: This algorithm is prone to overfitting with multicategory features, and its greedy algorithm may lead to local rather than global optimization. Additionally, deep trees can overfit training data, reducing their generalizability. RF: While RF integrates multiple DTs to enhance accuracy, it incurs high computational costs during training and testing, sacrificing the interpretability of individual decision trees. KELM: KELM is highly sensitive to kernel and regularization parameter selection and requires fine-tuning for optimal performance. Additionally, KELM, which is based on the kernel method, has a complex internal mechanism and lacks intuitive interpretability.

(2) Discussion on the limitations of deep learning methods

LeNet-5: As one of the earliest CNNs, it has a simple structure, leading to poor diagnostic performance under variable conditions and limited applicability in practical engineering scenarios. CNN: This method requires a large number of labeled datasets, with significant performance degradation when labeled data are limited.

CNN-LSTM, ACNN, WDCNN, GAN-CNN, HCAE, and GAMCAE are derivatives of CNNs, each with its shortcomings. In addition to the common requirement for a large number of high-quality labeled samples, specific issues include the following:

CNN-LSTM: Training the CNN and LSTM simultaneously. ACNN: Increased computational complexity and reduced model fitting efficiency due to the attention mechanism. WDCNN: The use of wide kernel convolution increases the number of model parameters, resulting in a more complex model that is more prone to overfitting on limited labeled datasets. GAN-CNN: Introducing the dynamic game process of

a GAN (generator and discriminator) makes the model prone to gradient vanishing and exploding, leading to training difficulties. Additionally, the generator may produce similar samples repeatedly, lacking data diversity and causing model collapse. It is also challenging to balance the generator and discriminator, often resulting in one overpowering the other and hindering effective learning. HCAE: This neural network combines data reconstruction and fault classification using multilayer convolution and transposed convolution, resulting in a high number of model parameters and increased time complexity. Additionally, the simultaneous optimization of reconstruction and classification objectives complicates model evaluation and training, increasing the likelihood of gradient vanishing and exploding. GAMCAE: This model combines a multiscale autoencoder with a GAN and faces similar challenges as other CNN and GAN-related models, such as a large number of parameters, training difficulties, gradient vanishing and exploding, and complexities in model evaluation.

(3) Advantages of WDSC-Net

WDSC-Net effectively uses limited labeled samples and is adaptable. By optimizing the wide kernel convolutional layer, 1×1 convolution, multiscale feature module based on DSC, and soft threshold denoising module, a simple, clear framework with low computational complexity is achieved. It has strong generalizability and adaptability in scenarios with limited labeled samples, variable conditions, and strong background noise.

Improved noise immunity. The enhanced soft thresholding method combined with deep learning reduces noise components in the signal, highlights important fault signal features, minimizes the complexity of manual threshold setting, and improves the noise immunity of the model. Optimization of computational resources. WDSC-Net significantly reduces the number of model parameters through multiscale channel convolution and point-by-point convolution of DSCs. Additionally, using 1×1 convolution and regularization methods, it designs a more efficient architecture that handles complex data of multiple fault types and avoids overfitting. Importantly, compared with other fault diagnostic methods, WDSC-Net reduces computational resource consumption and improves convergence speed, making it easier to deploy in real-world engineering scenarios.

5.2.4 Analysis of the computational complexity and resource requirements

WDSC-Net consists of five main modules: a wide kernel convolution module with a kernel size of 1 × 200, followed by a 1×1 convolution module for feature compression and extraction. The next module is the LMSFM, which uses multiscale depth-separable convolutions with kernel sizes of 1×3, 1×5, 1×7, and 1×9. Two LMSFM modules are designed in this paper, with channel numbers of 64 and 128. Finally, an improved soft-thresholding noise reduction module enhances feature extraction and noise reduction capabilities.

Table 5. Experimental results for comparative models.

Models	Accuracy/%	Test time/s	FLOPs/M	Params/K
Proposed	100	0.23	6.4	189
Model B[44]	98.40	0.4	39.7	334.4
Model C[45]	98.44	0.16	33.4	8499.2
Model D	98.75	0.43	80.1	198.5
Model E	96.25	0.27	121	7680
Model F	95.00	0.29	40.2	293.6
Model G[46]	95.63	0.28	49.5	20172.8
Model K	31.88	0.18	13.9	7168
Model M	39.38	42	4.2	2150
Model N	81.88	0.77	5.6	2867.2

As shown in Table 5, the method proposed in this paper excels in accuracy, test time, computational complexity, and parameters, making it ideal for deployment in resource-limited engineering scenarios. Specifically, the accuracy of the proposed method reaches 100%, the highest among all models. The computational complexity (FLOPs) is only 6.4 M,

indicating minimal computation. The model has 189 K parameters, indicating a low memory footprint. The inference time is 0.23 s, demonstrating fast inference. While other models may excel in certain aspects, none outperform the proposed method overall. For example, Model C is slightly less accurate but excels in test time and FLOPs, making it suitable for scenarios requiring high inference efficiency. In contrast, Model G is slightly more accurate but is suitable for scenarios with ample computational resources and high accuracy requirements due to its high parameter count and FLOPs. Low-performance models (e.g., Model K and Model M) perform poorly in terms of accuracy, computational complexity, and parameters, making them unsuitable for practical applications. In summary, our approach is efficient and well-suited for deployment in resource-constrained engineering scenarios.

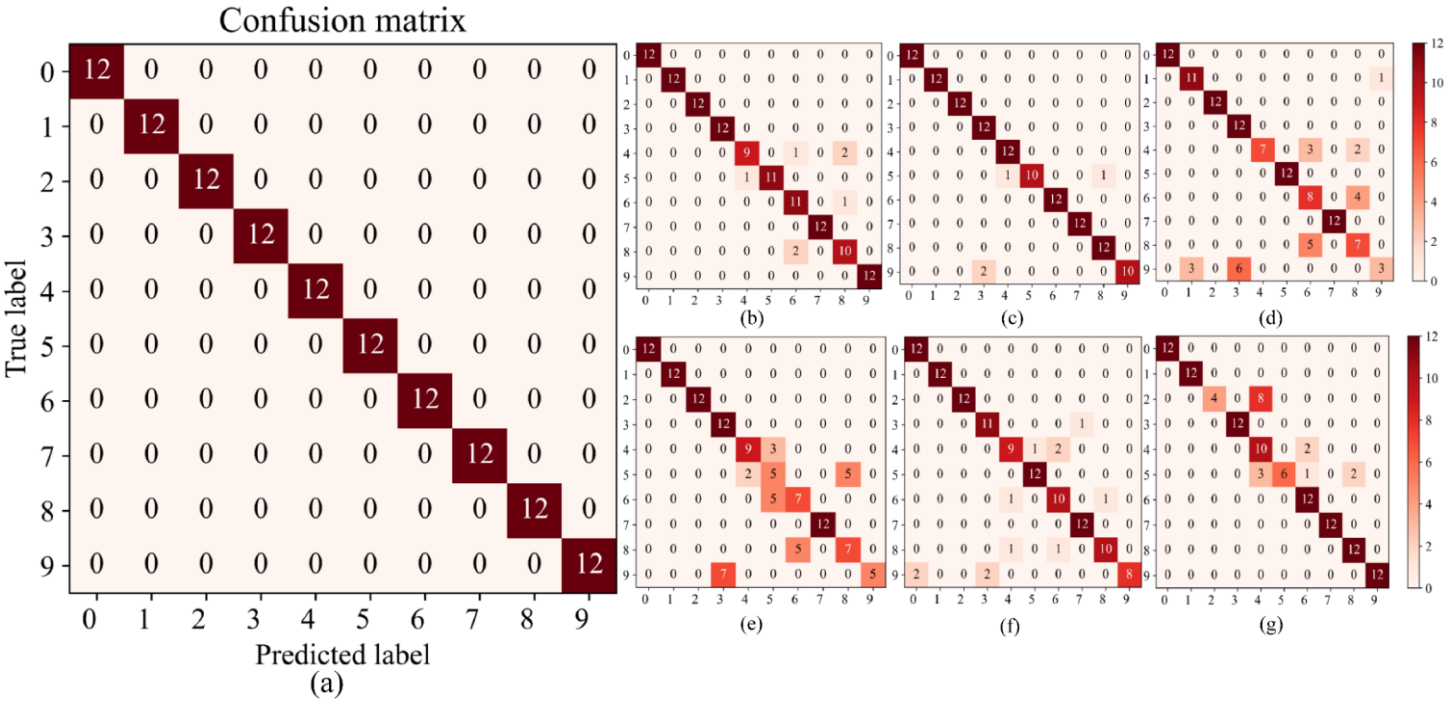


Fig. 12. Confusion matrix of test results of different models on dataset F1 (a) WDSC-Net, (b) ACNN, (c) WDCNN, (d) ResNet, (d) CNN-LSTM, (f) CNN, and (g) Let-5.

To further verify the superiority of the WDSC-Net method in identifying fault types, a confusion matrix was introduced to quantitatively analyze the identification results of Dataset F1. As shown in Fig. 12, only the method proposed in this study can distinguish fault types completely, whereas the other models exhibit varying degrees of misclassification. In particular, the ResNet, CNN-LSTM, and Let-5 models showed significant misclassification for bearing inner and outer ring faults, hindering their ability to perform the fault classification task effectively.

In summary, the method proposed in this study demonstrates a robust ability to recognize and classify various types of faults in rolling bearings, even when faced with limited labeled samples.

5.3 Comparison of the generalization performances of the proposed method and typical deep learning methods

In real-world industrial scenarios, the diagnostic capability of the model diminishes significantly owing to variations in workloads. Moreover, the likelihood of failure in normally operating machines is relatively low, and the available fault samples may be insufficient for the model to adapt effectively to changes. Hence, when faced with limited fault samples, it is imperative to assess diagnostic capability across different loads or speeds. To evaluate the adaptive capacity of the proposed method under varying loads and rotational speeds, a generalization performance experiment was conducted. Simultaneously, this study compares and analyses fault recognition models using six typical deep learning methods. Experimental validation was performed on the CWRU and

MCDS subdatasets, and the results are presented in Table 6.

Table 6. Recognition accuracies of different models under cross loading (CWRU subdataset).

Task	Proposed (A)	B [44]	C [45]	D	E	F	G [46]
F1→F2	96.61%	77.60%	81.25%	68.06%	53.30%	77.60%	68.93%
F1→F3	86.46%	79.17%	86.20%	66.67%	57.12%	79.17%	69.27%
F1→F4	86.28%	77.78%	80.47%	63.72%	51.04%	77.78%	59.72%
F2→F1	86.98%	82.99%	79.43%	75.69%	60.94%	78.30%	61.63%
F2→F3	97.57%	95.49%	94.53%	87.15%	76.56%	85.94%	82.99%
F2→F4	95.14%	94.62%	84.38%	73.10%	66.84%	78.30%	75.86%
F3→F1	94.79%	95.83%	86.20%	71.35%	47.92%	83.16%	64.58%
F3→F2	92.36%	86.46%	86.72%	85.24%	76.91%	87.15%	83.85%
F3→F4	98.70%	98.44%	95.57%	74.65%	73.44%	85.60%	89.24%
F4→F1	94.79%	94.79%	72.40%	78.30%	49.65%	81.42%	64.41%
F4→F2	97.92%	90.63%	76.04%	72.74%	72.22%	67.89%	76.56%
F4→F3	92.97%	92.36%	84.90%	77.08%	68.58%	80.56%	73.26%
Ave acc	93.38%	88.84%	84.00%	74.48%	62.88%	80.24%	72.53%

In cross-load validation experiments on the CWRU subdataset, the WDSC-Net model proposed in this study achieved the highest average recognition accuracy of 93.38% across 12 working conditions, outperforming the other six typical fault recognition models. Model B, the best performer among the 6 common models, achieves an average recognition accuracy of 88.84%, which is still 4.54% lower than that of the proposed model; notably, for working condition F3→F1, the recognition accuracy of Model B exceeds that of the proposed method by 1.04%. However, under all the other working conditions, the recognition accuracy of Model B falls short of that of the proposed method, and Model B shows greater potential in fault identification. Model E exhibits the poorest fault recognition performance among all the models, with values that are only 62.88%, 30.5% lower than those of Model A.

Table 7. Recognition accuracy of different models cross-load and variable-speed conditions.

Task	Proposed (A)	B [44]	C [45]	D	E	F	G [46]
F5→F6	97.50%	81.25%	54.95%	60.94%	49.48%	59.38%	62.50%
F5→F7	75.00%	51.25%	42.75%	50.00%	26.04%	37.50%	26.00%
F5→F8	67.50%	43.75%	25.00%	45.50%	25.00%	35.94%	21.88%
F6→F5	72.92%	61.25%	67.00%	40.00%	61.88%	37.50%	32.81%
F6→F7	90.00%	77.50%	46.88%	75.25%	40.63%	84.38%	21.88%
F6→F8	93.75%	68.75%	30.63%	66.75%	27.08%	40.63%	21.88%
F7→F5	72.50%	48.30%	47.50%	27.08%	42.97%	31.25%	32.81%
F7→F6	76.25%	57.91%	68.49%	41.40%	48.25%	48.44%	42.19%
F7→F8	100.00%	90.00%	78.65%	75.52%	35.00%	84.38%	62.50%
F8→F5	44.25%	36.88%	25.00%	26.04%	25.00%	31.25%	25.00%
F8→F6	49.58%	36.25%	31.25%	37.50%	31.25%	33.75%	31.25%
F8→F7	71.25%	70.73%	40.37%	54.75%	54.68%	72.50%	40.63%
Ave Acc	75.88%	60.32%	46.54%	50.06%	38.94%	49.74%	35.11%

According to the results of cross-speed validation experiments on the MCDS subdataset, the recognition accuracy

of the proposed WDSC-Net model was the highest among the six fault recognition models across the 12 cross-speed conditions. Model B continued to exhibit the best classification performance among the six typical recognition models, but it was lower than that of the WDSC-Net model. Comparing the fault classification abilities of models A and B under 12 working conditions, model A outperformed model B under all conditions. Moreover, the average accuracy of Model G under 12 working conditions was only 35.11%, indicating that it performed the worst among all the models.

To further validate the generalization performance of the model in noisy scenarios, we conducted two tasks: F2-F3 and F5-F6. F2-F3 refers to training the model with clean samples in F2 while injecting noise with different signal-to-noise ratios (SNR) into the test samples in F3. Similarly, F5-F6 involves training the model in F5 and testing it in F6.

As shown in Table 8, the proposed method performs optimally at all SNR levels in both the F2-F3 and F5-F6 tasks, particularly at higher SNR levels. This indicates that the proposed method has significant advantages and strong generalizability for handling noise and variable conditions. Specifically, Model B and Model F perform better at high SNR levels but worse at low SNR levels, indicating weaker generalization abilities. Model C performs stably at all SNR levels, but its overall generalization performance is slightly lower than that of the proposed method. In contrast, Model E and Model G show poor diagnostic performance in both tasks and have the weakest generalization abilities. In summary, the proposed method performs well under variable conditions and noisy scenarios, maintaining stable diagnostic performance across all SNR levels, reflecting its strong generalizability.

Table 8. Generalization performance analysis of different models in noisy environments.

Methods	Tasks	SNR(dB)						Ave acc
		-4	-2	0	2	4	6	(%)
Proposed (A)	F2- F3	75.90	77.10	77.20	77.40	78.40	87.80	78.97
Model B [44]	F2- F3	61.70	66.30	69.10	80.20	83.60	84.10	74.16
Model C [45]	F2- F3	66.30	70.00	71.00	75.00	76.40	81.80	73.41
Model D	F2- F3	54.80	58.70	65.90	66.50	68.70	73.90	64.75
Model E	F2- F3	57.90	58.50	59.90	60.00	61.90	64.60	60.46
Model F	F2- F3	68.40	71.20	72.60	74.50	76.70	77.40	73.46
Model G [46]	F2- F3	62.20	63.90	65.40	66.60	68.20	71.70	66.33
Ave acc (%)	/	63.89	66.53	68.73	71.46	73.41	77.33	/
Proposed (A)	F5-F6	76.00	76.25	78.75	84.50	86.25	90.75	82.08
Model B [44]	F5-F6	55.88	67.25	67.63	68.63	72.75	74.13	67.71
Model C [45]	F5-F6	57.88	60.38	69.00	69.50	72.13	74.50	67.23
Model D	F5-F6	71.75	72.13	75.38	79.13	82.50	83.63	77.42
Model E	F5-F6	35.63	41.25	41.50	45.88	47.25	47.38	43.14
Model F	F5-F6	61.63	62.50	64.00	68.38	71.00	81.63	68.19
Model G [46]	F5-F6	32.50	34.50	36.63	39.25	51.00	58.50	42.06
Ave acc (%)	/	55.90	59.34	61.84	65.04	68.98	72.93	/

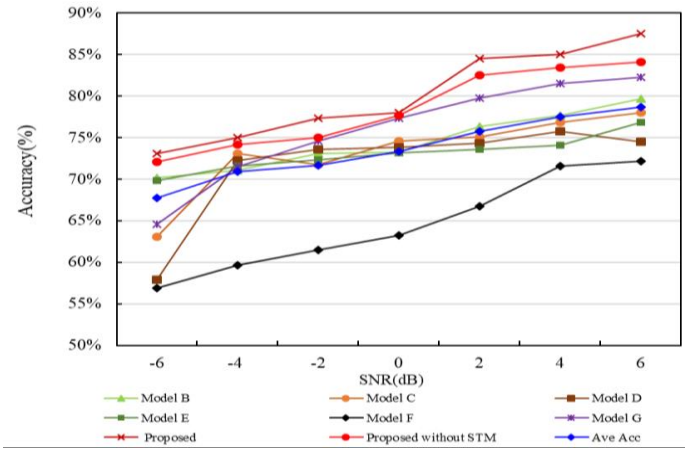
In summary, under limited labeled samples, the proposed WDSC-Net model demonstrated excellent adaptability to variable conditions and performed exceptionally well under both cross-load and variable-speed conditions.

5.4 The analysis of anti-noise performance

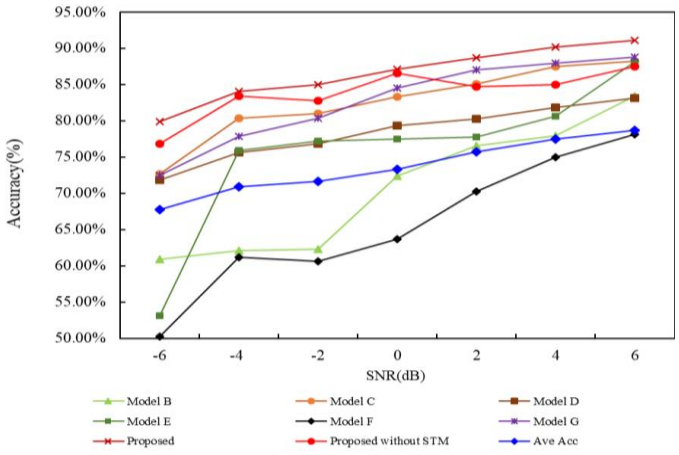
Owing to the complexity of the industrial environment, vibration signal acquisition is often subject to varying degrees of noise pollution, leading to the obscuring of fault information in the signal. To simulate a real industrial noise environment, synthetic noise was introduced to the original signal, enhancing the robustness of the model to noise. In this study, Gaussian noise with a signal-to-noise ratio of -6-6 dB was applied to the limited labeled samples in datasets F1 and F5 to emulate different intensities of noise in a real industrial setting. A comparative analysis was then conducted with six types of deep learning fault models to validate the antinoise performance of the proposed method.

Table 9. Recognition effects of different models in noisy environments.

Methods	Datasets	SNR (dB)						
		-6	-4	-2	0	2	4	6
Proposed (A)	F1	73.13%	75.00%	77.38%	78.00%	84.50%	85.00%	87.50%
A without STM	F1	72.07%	74.15%	75.00	77.73%	82.55%	83.46%	84.11
Model B [44]	F1	70.20%	71.13%	73.13%	73.25%	76.38%	77.73%	79.69%
Model C [45]	F1	63.13%	73.13%	71.75%	74.63%	75.13%	76.82%	78.00%
Model D	F1	57.88%	72.25%	73.63%	73.87%	74.37%	75.75%	74.5%
Model E	F1	69.88%	71.63%	72.38%	73.17%	73.57%	74.09%	76.82%
Model F	F1	56.88%	59.63%	61.50%	63.25%	66.80%	71.61%	72.20%
Model G [46]	F1	64.63%	71.50%	74.63%	77.38%	79.75%	81.5%	82.29%
Ave acc	F1	67.74%	70.95%	71.72%	73.36%	75.79%	77.50%	78.71%
Proposed (A)	F5	79.94%	84.06%	85.00%	87.18%	88.75%	90.23%	91.15%
A without STM	F5	76.88%	83.43%	82.81%	86.56%	84.69%	85.00%	87.50%
Model B [44]	F5	60.93%	62.10%	62.28%	72.43%	76.56%	78.00%	83.46%
Model C [45]	F5	72.69%	80.42%	81.04%	83.33%	85.15%	87.50%	88.28%
Model D	F5	71.87%	75.63%	76.88%	79.38%	80.31%	81.88%	83.20%
Model E	F5	53.13%	75.94%	77.19%	77.50%	77.81%	80.63%	88.13%
Model F	F5	50.31%	61.25%	60.63%	63.75%	70.31%	75.00%	78.15%
Model G [46]	F5	72.50%	77.91%	80.42%	84.58%	87.08%	87.94%	88.84%
Ave acc	F5	63.78%	73.90%	74.78%	78.31%	80.85%	83.03%	85.88%



(a) Dataset F1



(b) Dataset F5

Fig. 13 Accuracy of various methods on datasets F1 and F5 under strong noise

As shown in Table 9 and Fig. 13(a), with the introduction of the soft threshold (STM) feature denoising module, the proposed method has the best fault identification performance in noisy environments with signal-to-noise ratios ranging from -6 to 6 dB, which is superior to the average accuracies of all methods by 4%-8%. Specifically, the diagnostic accuracies of the WDSC-Net model in strong noise environments are 73.13% and 87.5%, respectively. When STM is not introduced, the diagnostic performance of the proposed model in a strong noise environment decreases, but it is still superior to other typical comparison methods, and it also illustrates the superior feature extraction capability of the proposed method in this paper. The fault identification capability of each model improves as the signal-to-noise ratio increases. At a signal-to-noise ratio of 6 dB, the recognition accuracy of model G is the highest among the six diagnostic models except for the method proposed in this paper, but it is still 5.21% lower than the accuracy of the proposed model.

According to Table 9 and Fig. 13(b), the proposed method exhibits the best recognition performance when the STM is

introduced. In a strong noise environment with a signal-to-noise ratio of -6 dB, Model F has the lowest recognition accuracy of 50.31%. The recognition accuracies of Model C, Model D, and Model G are close to that of the proposed method, but still approximately 8% lower than that of the proposed method WDSC-Net, which illustrates that the method in this paper can effectively remove the noise from the signal while retaining the effective components and continuity of the signal in the lower signal-to-noise ratio scenario. When the signal-to-noise ratio is increased to 4 dB and 6 dB, the proposed method can achieve 90% recognition performance. When STM is not introduced, the recognition performance is the best among all methods when the signal-to-noise ratios are -6, -4, -2, and 0, which illustrates the powerful feature extraction capability of our method.

In conclusion, the proposed WDSC-Net recognition model demonstrates outstanding defect recognition capability even in diverse noisy environments with a limited number of labeled samples.

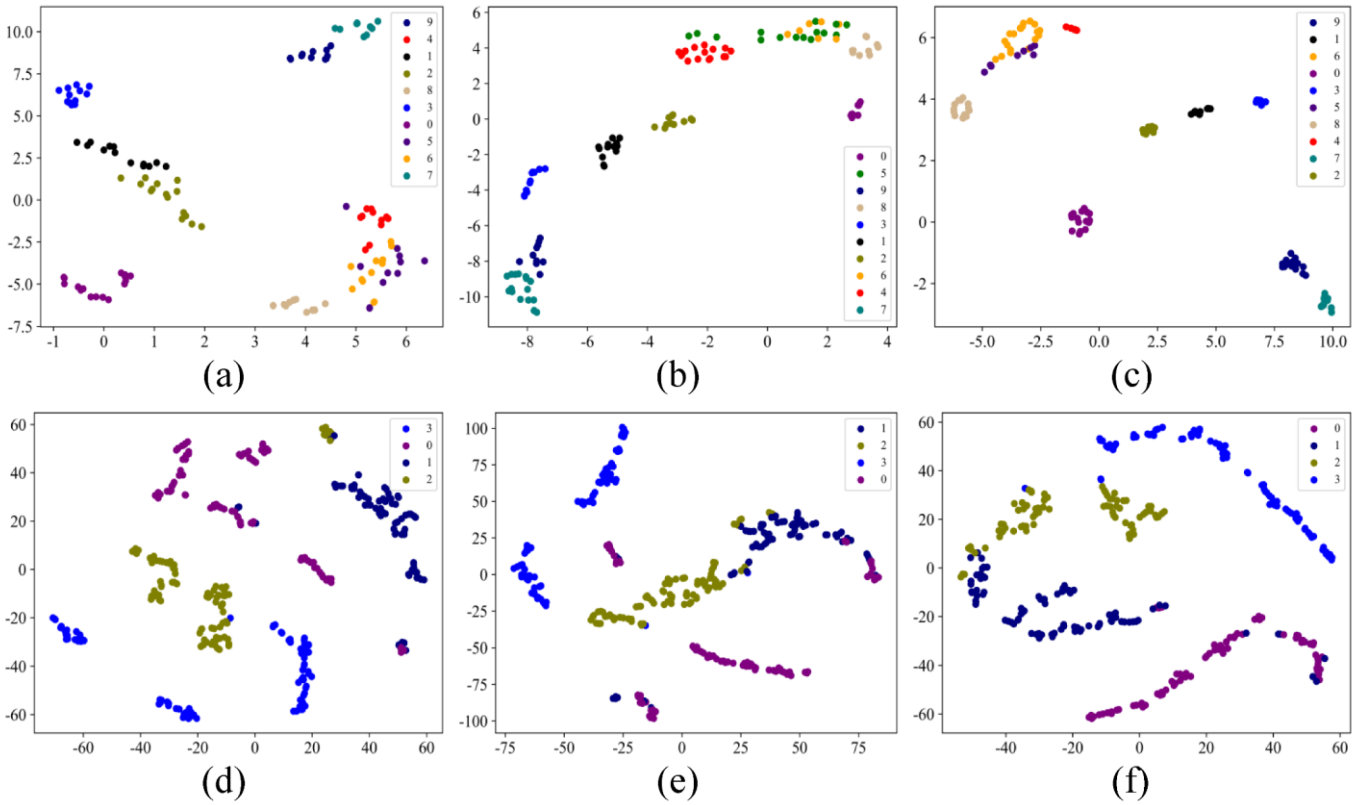
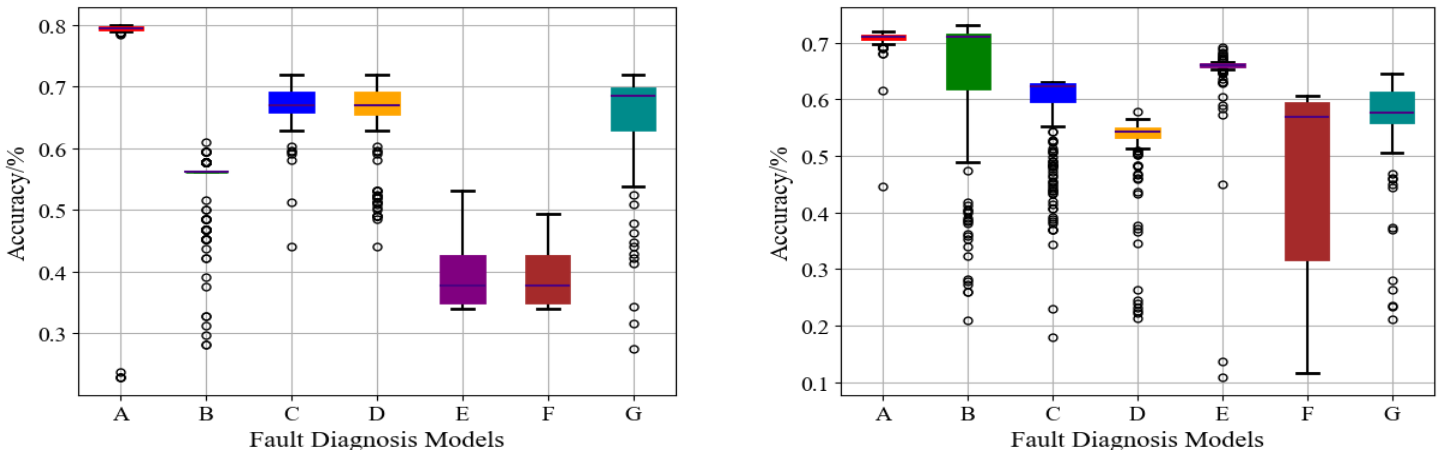


Fig. 14. T-SNE of WDSC-Net with different numbers of layers when the SNR=6 dB.



(a) Dataset F1 (b) Dataset F5

Fig. 15. Recognition performance of various models in an environment with an SNR of -6 dB.

To gain a comprehensive understanding of the network layer training process, t-stochastic neighbor embedding (t-SNE) was utilized with an SNR of 6 dB in the test sets of datasets F1 and F5. Fig.14 shows the classification results for the wide-kernel convolutional, multiscale feature fusion, and FC layers. Fig.14 also displays the distribution of fault class in different noise-intensity environments. After a wide kernel convolution, the state data begin to separate with clear boundaries. Subsequently, the multiscale feature fusion layer further reduced feature overlap. With the FC layer, the bearing health states can be accurately differentiated, resulting in a narrowing of the space between different classes and a widening of the space between categories. Clear boundaries emerge between samples of different categories, confirming the robustness of the proposed model in terms of fault classification ability in strong noise environments. The proposed model consistently exhibited stable performance in fault classification under strong noise conditions.

To further visualize the recognition performance of the WDSC-Net model under strong noise interference, box-and-line plots were generated to illustrate the stability of the recognition results among the different models. The experiments were performed in a noisy environment with an SNR of -6 dB for the F1 and F5 datasets. As depicted in Fig. 15, the WDSC-Net model proposed in this study yields superior recognition results with a smaller variance in the strong-noise environment of dataset F1. Except for models B and E, which show smaller variances in recognition accuracy in datasets F1 and F5, respectively, the recognition results of the other models are not sufficiently stable in strong noise environments, and there are a significant number of outliers.

In summary, under the constraint of limited labeled samples, the WDSC-Net model proposed in this study significantly enhances the robustness of fault recognition ability in noisy environments. Additionally, the model demonstrated better adaptability to different bearings and fault types.

5.5 Ablation experiment

To intuitively analyze the impact of the main modules on the diagnostic performance of the proposed model, we constructed WDSC-Net using the wide kernel convolution module (WKC), LMSFM, and soft threshold denoising module (STM). We conducted ablation experiments on a self-constructed bearing dataset. Two hundred samples were taken from each class, and to ensure the reliability of the results, each experiment was repeated three times. The mean values were taken as the final results. The experiments were evaluated based on model diagnostic accuracy, the number of model parameters, and computational complexity (FLOPs). The specific experimental results are shown in Table 10.

Table 10. Ablation experiments on a self-built bearing dataset.

WKC	LMSFM1	LMSFM2	STM	Accuracy（%）	Params/K	FLOPs/M
√			√	25.83	3.3	6.0
	√		√	96.50	38.4	10.9
		√	√	96.40	142.3	37.6
√	√		√	96.70	38.6	5.3
√		√	√	97.20	142.5	17.6
√	√	√	√	99.93	189	6.4

As shown in Table 10, different module combinations affect the classification accuracy and computational resources as follows:

Using only the WKC and STM modules results in an accuracy of only 25.83%. This suggests that their feature extraction capability is insufficient, likely because the features captured by WKC are too coarse and the denoising effect is limited. In this scenario, the model has few parameters and low computational requirements, indicating low complexity but poor classification performance. Combining the LMSFM1 and STM modules significantly increases the accuracy to 96.50%. This indicates that this combination can effectively extract multiscale features and substantially improve classification performance. In this scenario, the number of parameters and computational effort are moderate, suggesting that the complexity and computational requirements are within

acceptable limits. The accuracy is slightly lower with the combination of the LMSFM2 and STM modules than with the LMSFM1 and STM modules, and the improvement is not significant. This may be due to the greater complexity of the LMSFM2, which did not significantly enhance the performance in this experiment despite its ability to extract effective features. Combining the WKC, LMSFM1, and STM modules yields an accuracy of 96.70%, which is slightly higher than that of previous combinations. This indicates that the WKC, LMSFM1, and STM algorithms can better extract features with a moderate number of parameters and fewer computations, striking a balance between performance and computational efficiency. Combining the WKC, LMSFM2, and STM modules improves the accuracy to 97.20%. This suggests that WKC, LMSFM1, and LMSFM2 can fully utilize multiscale and wide kernel features for effective classification, further enhancing accuracy. However, this combination also increases the number of parameters and computations.

Combining the WKC, LMSFM1, LMSFM2, and STM modules results in an accuracy of 99.93%, the highest value observed. This demonstrates that this combination can extract the most effective features and significantly improve classification performance. Despite a significant increase in the number of parameters, the computational load remains low, indicating high computational efficiency while maintaining high performance. In summary, the method proposed in this paper achieves an optimal balance between diagnostic accuracy, model parameters, and computational complexity.

6. Conclusion and future work

We introduce WDSC-Net, a novel, lightweight, and efficient deep learning model for end-to-end rotating machinery fault diagnosis. This model outperforms several state-of-the-art methods by utilizing standard convolutions while significantly reducing the number of model parameters, making it suitable for environments with a limited number of labeled samples. The main contributions and conclusions of this study are summarized as follows:

(1) In the context of fault localization in rotating machinery and the variability of response times for different faults, relatively small convolution kernels may not adequately represent the full range of information regarding the fault impact. Therefore, in this study, a wide kernel convolution was employed to expand the perceptual field of view, extract more global information about different states, and localize the fault-affected segment.

(2) WDSC-Net is a lightweight model in terms of storage and computational efficiency. It incorporates a multiscale feature fusion network constructed based on DSC. This construction enabled the model to generate differentiated features with mixed spatial location information. Cross-channel correlations were mapped using point-by-point convolutional mapping of the DSC. This design allowed the model to concentrate on more differentiated features at various locations in the network. Convolutional kernels of different sizes were intentionally designed to identify the most relevant feature mapping in faults. This approach helps the model learn the underlying relationships between the inputs and outputs of the network, resulting in highly discriminative features.

(3) To alleviate the impact of noise on diagnostic accuracy in real-world engineering scenarios, this study introduces a new approach that combines the soft thresholding method with deep learning. By integrating these techniques, this study achieved recognition accuracies of 73.13% and 79.94% on the CWRU and MCDS datasets, respectively, under a signal-to-noise ratio of -6 dB. This innovative approach not only provides a new perspective for solving this problem but also eliminates the tedious and subjective task of manually setting thresholds.

(4) We evaluate the generalization performance of the WDSC-Net model with limited labeled samples. Twelve types of scenarios are designed on the CWRU and MCDS limited-labeled fault sample datasets respectively, and the results show that the proposed model has excellent diagnostic performance in various scenarios under variable working conditions, with average accuracies of 93.38% and 75.88%, respectively, which are better than those of other typical deep learning fault diagnosis methods.

(5) In future work, we will aim to enhance the generalization ability and robustness of the WDSC-Net model. Simultaneously, given the varied probability distributions of the collected data and the limited availability of test samples, we plan to undertake transfer learning under the constraint of limited labeled samples. This approach was intended to broaden the applicability of the proposed approach in the field of engineering.

Acknowledgments

This work was partly supported by the National Key Research and Development Program of China under Grant 2019YFB1704500, partly by the State Ministry of Science and Technology Innovation Fund of China under Grant 2018IM030200, the National Natural Foundation of China under Grant U1708255, and in part by the National Science and Technology Major Project under Grant J2019-V-0009-0103.

References

1. Zhou H, Huang X, Wen G et al. Construction of health indicators for condition monitoring of rotating machinery: A review of the research. Expert systems with applications 2022; 203 117297, https://doi.org/10.1016/j.eswa.2022.117297.
2. Li Z, Chen J, Zi Y et al. Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive. Mechanical systems and signal processing 2017; 85 512-529, https://doi.org/10.1016/j.ymssp.2016.08.042.
3. Zhang J, Kong X, Cheng L et al. Intelligent fault diagnosis of rolling bearings based on continuous wavelet transform-multiscale feature fusion and improved channel attention mechanism. Eksploatacja i Niezawodność, 2023, 25(1), http://doi.org/10.17531/ein.2023.1.16.
4. Kapoor R, Gupta R, Jha S et al. Boosting performance of power quality event identification with KL Divergence measure and standard deviation. Measurement 2018; 126 134-142, https://doi.org/10.1016/j.measurement.2018.05.053.
5. Zhang Y, Xing K, Bai R et al. An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image. Measurement 2020; 157 107667, https://doi.org/10.1016/j.measurement.2020.107667.
6. Zhao R, Yan R, Chen Z et al. Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 2019, 115: 213-237, https://doi.org/10.1016/j.ymssp.2018.05.050.
7. Gong W, Wang Y, Zhang M et al. A fast anomaly diagnosis approach based on modified CNN and multisensor data fusion. IEEE Transactions on Industrial Electronics 2021; 69(12) 13636-13646, https://doi.org/10.1109/TIE.2021.3135520.
8. Xu G, Liu M, Jiang Z et al. Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. Sensors 2019; 19(5) 1088, https://doi.org/10.3390/s19051088.
9. Fuan W, Hongkai J, Haidong S et al. An adaptive deep convolutional neural network for rolling bearing fault diagnosis. Measurement Science and Technology 2017; 28(9) 095005, https://doi.org/10.1088/1361-6501/aa6e22.
10. Yang J, Zhang Y, Zhu Y. Intelligent fault diagnosis of rolling element bearing based on SVMs and fractal dimension. Mechanical Systems and Signal Processing 2007; 21(5) 2012-2024, https://doi.org/10.1016/j.ymssp.2006.10.005.
11. Ye Z, Yu J. 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, https://doi.org/10.1016/j.ymssp.2021.107984.
12. Ke H, Chen D, Li X et al. Towards brain big data classification: Epileptic EEG identification with a lightweight VGGNet on global MIC. IEEE Access 2018; 6 14722-14733, https://doi.org/10.1109/ACCESS.2018.2810882.
13. Lu S, Lu Z, Zhang Y D. Pathological brain detection based on AlexNet and transfer learning. Journal of computational science 2019; 30 41-47, https://doi.org/10.1016/j.jocs.2018.11.008.
14. Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern recognition 2019; 90 119-133, https://doi.org/10.1016/j.patcog.2019.01.006.
15. Fang H, Deng J, Zhao B et al. LEFE-net: A lightweight efficient feature extraction network with strong robustness for bearing fault diagnosis. IEEE Transactions on Instrumentation and Measurement 2021; 70: 1–11, https://doi.org/10.1109/TIM.2021.3067187.
16. Deng J, Jiang W, Zhang Y et al. HS-KDNet: A lightweight network based on hierarchical-split block and knowledge distillation

for fault diagnosis with extremely imbalanced data. IEEE Transactions on Instrumentation and Measurement 2021; 70: 1–9, https://doi.org/10.1109/TIM.2021.3091498.

17. Lu R, Liu S, Gong Z et al. Lightweight Knowledge Distillation-Based Transfer Learning Framework for Rolling Bearing Fault Diagnosis. Sensors 2024; 24(6): 1758, https://doi.org/10.3390/s24061758.
18. Xiong S, Zhou H, He S et al. Fault diagnosis of a rolling bearing based on the wavelet packet transform and a deep residual network with lightweight multi-branch structure. Measurement Science and Technology 2021; 32: 085106, https://doi.org/10.1088/1361-6501/abe448
19. Liu W, Guo P, Ye L. A low-delay lightweight recurrent neural network (LLRNN) for rotating machinery fault diagnosis. Sensors 2019; 19: 3109, https://doi.org/10.3390/s19143109.
20. Cui J, Zhong Q, Zheng S et al. A lightweight model for bearing fault diagnosis based on gramian angular field and coordinate attention. Machines 2022; 10: 282, https://doi.org/10.3390/machines10040282.
21. Qin Z, Li Z, Zhang Z, et al. ThunderNet: Towards real-time generic object detection on mobile devices. Proceedings of the IEEE/CVF international conference on computer vision. 2019; 6718-6727. https://doi.org/10.1109/ICCV.2019.00682
22. Iandola FN, Han S, Moskewicz MW et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv 2016; https://doi.org/10.48550/arXiv.1602.07360.
23. Li Z Q. Adaptive Anti-noise Fault Diagnosis Algorithm for Bearings based on An Improved Convolution Neural Network. Machinery Design & Manufacture 2024;1-7. https://doi.org/10.19356/j.cnki.1001-3997.20240321.015.
24. Dong R, Xu Y W, Long Z H et al. Fault Diagnosis Model of Noise-resistant Bearings Using Large Kernel Attention Mechanism. Noise and Vibration Control, 2023; 43(02): 162-168, https://doi.org/10.3969/j.issn.1006-1355.2023.02.024.
25. Li R, Wu J, Li Y et al. Periodnet: Noise-robust fault diagnosis method under varying speed conditions. IEEE Transactions on Neural Networks and Learning Systems 2023; https://doi.org/ 10.1109/TNNLS.2023.3274290
26. Fan W, Chen Z, Li Y et al. A reinforced noise resistant correlation method for bearing condition monitoring. IEEE Transactions on Automation Science and Engineering 2022; 20(2): 995-1006, https://doi.org/10.1109/TASE.2022.3177010.
27. Dong J, Jiang H, Su D et al. Transfer learning rolling bearing fault diagnosis model based on deep feature decomposition and class-level alignment. Measurement Science and Technology, 2024; 35(4): 046006, https://doi.org/10.1088/1361-6501/ad2052.
28. Yang G, Liu L, Xi C. Bearing fault diagnosis based on sa-acgan data generation model. China Mechanical Engineering 2022; 33(13) 1613, https://doi.org/10.3969/j.issn.1004-132X.2022.13.012.
29. Meng A, Chen S, Ou Z et al. A novel few-shot learning approach for wind power prediction applying secondary evolutionary generative adversarial network. Energy 2022; 261: 125276, https://doi.org/10.1016/j.energy.2022.125276.
30. Lei Y, Yang B, Jiang X et al. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing 2020; 138 106587, https://doi.org/10.1016/j.ymssp.2019.106587.
31. Meng X, Hu T, Li J et al. A digital twin library of mechanical transmission system for the application of small sample fault diagnosis problem. Measurement Science and Technology 2024; 35(6): 066125, https://doi.org/10.1088/1361-6501/ad34ef.
32. Yang J, Liu J, Xie J et al. Conditional GAN and 2-D CNN for bearing fault diagnosis with small samples. IEEE Transactions on Instrumentation and Measurement 2021; 70: 1-12, https://doi.org/10.1109/TIM.2021.3119135.
33. Zhou X K, Yu J B. Gearbox Fault Diagnosis Based on One-dimension Residual Convolutional Auto-encoder. Journal of Mechanical Engineering 2020; 56(7), 96-108. (in Chinese). https://doi.org/10.3901/JME.2020.07.096
34. Wang H, Liu Z, Peng D et al. Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis. IEEE Transactions on Industrial Informatics 2019; 16(9) 5735-5745, https://doi.org/10.1109/TII.2019.2955540.
35. He Z, Yang L, Angizi S et al. Sparse BD-Net: A multiplication-less DNN with sparse binarized depth-wise separable convolution.

ACM Journal on Emerging Technologies in Computing Systems (JETC) 2020; 16(2) 1-24, https://doi.org/10.1145/3369391.

36. Zhang J, Kong X, Cheng L et al. Intelligent fault diagnosis of rolling bearings based on continuous wavelet transform-multiscale feature fusion and improved channel attention mechanism. Eksploatacja i Niezawodność 2023; 25(1), http://doi.org/10.17531/ein.2023.1.16.
37. Zhang J, Xiangwei K, Xueyi L I et al. Fault diagnosis of bearings based on deep separable convolutional neural network and spatial dropout. Chinese Journal of Aeronautics 2022; 35(10) 301-312, https://doi.org/10.1016/j.cja.2022.03.007.
38. Xu X, Wang J, Zhong B et al. Deep learning-based tool wear prediction and its application for machining process using multi-scale feature fusion and channel attention mechanism. Measurement 2021; 177 109254, https://doi.org/10.1016/j.measurement.2021.109254.
39. Sun Y, Weng Y, Luo B et al. Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. IET Image Processing 2023; 17(4) 1280-1290, https://doi.org/10.1049/ipr2.12712.
40. Guo J, Si Z, Xiang J. A compound fault diagnosis method of rolling bearing based on wavelet scattering transform and improved soft threshold denoising algorithm. Measurement 2022; 196: 111276, https://doi.org/10.1016/j.measurement.2022.111276.
41. Zhao M, Zhong S, Fu X et al. Deep residual shrinkage networks for fault diagnosis. IEEE Transactions on Industrial Informatics 2019; 16(7) 4681-4690, https://doi.org/ 10.1109/TII.2019.2943898.
42. Loparo, K A. A bearing data center. Retrieved from http://www.eecs.case.edu/laboratory/bearing/welcome_overview.htm.
43. Li X, Zhang W. Deep learning-based partial domain adaptation method on intelligent machinery fault diagnostics. IEEE Transactions on Industrial Electronics 2020; 68(5) 4351-4361; https://doi.org/10.1109/TIE.2020.2984968.
44. Sun L, Zhu X, Xiao J et al. A hybrid fault diagnosis method for rolling bearings based on GGRU-1DCNN with AdaBN algorithm under multiple load conditions. Measurement Science and Technology 2024; 35(7): 076201, https://doi.org/10.1088/1361-6501/ad3669.
45. Zhang W, Peng G, Li C 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, https://doi.org/10.3390/s17020425.
46. Zhu Y, Li G, Wang R et al. Intelligent fault diagnosis of hydraulic piston pump combining improved LeNet-5 and PSO hyperparameter optimization. Applied Acoustics 2021; 183: 108336, https://doi.org/10.1016/j.apacoust.2021.108336.
47. Xu H, Wamg Z. Condition Evaluation and Fault Diagnosis of Power Transformer Based on GAN-CNN. Journal of Electrotechnology, Electrical Engineering and Management 2023; 6(3) 8-16, https://doi.org/10.23977/jeeem.2023.060302.
48. Hu Z, Han T, Bian J et al. A deep feature extraction approach for bearing fault diagnosis based on multi-scale convolutional autoencoder and generative adversarial networks. Measurement Science and Technology 2022; 33(6) 065013, https://doi.org/10.1088/1361-6501/ac56f0.
49. Wu X, Zhang Y, Cheng C et al. A hybrid classification autoencoder for semi-supervised fault diagnosis in rotating machinery. Mechanical Systems and Signal Processing 2021; 149 107327, https://doi.org/10.1016/j.ymssp.2020.107327.