A lightweight mechanical fault diagnosis framework based on dynamic separable convolution and broadcast self-attention

Peiyi Yang[1], Yizhe Song[2], Tiantong Zhang[3]
[1, 3]College of Locomotive and Rolling Stock, Zhengzhou Railway Vocational and Technical College, Zhengzhou, 450046, China
[2]School of Mechatronics and Vehicle Engineering, East China Jiaotong University, Nanchang, 330013, China
[3]Corresponding author
E-mail: [1]*yang_peiyi123@163.com*, [2]*15095042867@163.com*, [3]*405919311@qq.com*

Received 11 March 2026; accepted 7 June 2026; published online 19 August 2026
DOI https://doi.org/10.21595/jve.2026.26313

Check for updates

Copyright © 2026 Peiyi Yang, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. To address issues such as the large number of parameters, high computational complexity, and inadequate real-time performance in existing CNN-Transformer hybrid fault diagnosis models, we propose a lightweight fault diagnosis framework, LWConvFormer. This framework comprises two core innovative modules: a dynamic separable multi-scale convolutional module that employs a gated network for adaptive feature extraction, thereby reducing computational load while enhancing adaptability to complex fault modes; and a broadcast self-attention module that substitutes traditional matrix multiplication with broadcast operations, thereby decreasing computational complexity from a quadratic to a linear level. Experimental results based on the planetary gearbox at Xi'an Jiaotong University and the QPZZ-II type rotating machinery test bench demonstrate that LWConvFormer maintains excellent diagnostic performance across various noise levels. The number of parameters and computational load are reduced by a factor of 6 to 10 compared to mainstream methods, while the training speed increases by nearly 7 times. This framework effectively balances diagnostic accuracy, model lightweighting, and noise resistance, offering an efficient solution for real-time fault diagnosis in industrial settings.

Keywords: lightweight fault diagnosis, dynamic separable multi-scale convolution, broadcast self-attention, noise robustness.

1. Introduction

The health status of rotating machinery, as essential equipment in the industrial sector, is critical for ensuring safe operations and minimizing maintenance expenses [1]. Prolonged exposure to high-intensity working environments leads mechanical systems to encounter various failure modes. Consequently, the advancement of sophisticated fault diagnosis technologies has emerged as a significant research focus within the framework of Industry 4.0 and intelligent manufacturing [2].

Traditional fault diagnosis methods depend on expert experience and signal processing techniques, which exhibit considerable limitations. In contrast, deep learning, characterized by its robust automatic feature extraction capabilities, has emerged as a pivotal approach for advancing intelligent fault diagnosis [3]. Notably, Convolutional Neural Networks (CNNs) and Transformer models have demonstrated exceptional performance [4]. The current methods encounter significant challenges. Traditional CNN models depend on cross-channel convolution mechanisms, leading to excessively high model complexity. Although the multi-head self-attention mechanism of the Transformer can capture global features, its secondary computational complexity constrains its application in industrial settings [5]. In recent years, researchers have sought to enhance diagnostic models through various strategies. Tang et al. [6]

developed a CNN framework incorporating normalization strategies for fault diagnosis in hydraulic systems; however, CNN performance is limited in environments with strong noise. The Transformer model, recognized for its robust sequence modeling capability, has gradually found applications in fault diagnosis [7]. Nonetheless, it lacks inductive bias and necessitates substantial training data. Consequently, the CNN-Transformer hybrid architecture has emerged as a promising research direction, enabling the simultaneous extraction of local features and global dependencies [8]. Despite this, existing hybrid architectures exhibit notable deficiencies: the computational load of the embedded CNN module remains relatively high, while the Transformer module requires complex matrix operations, making it challenging to satisfy real-time requirements [9]. Although some studies have achieved lightweight models through feature compression, this approach compromises fine-grained feature information [10].

In response to the above problems, this paper proposes a lightweight fault diagnosis framework, LWConvFormer. The main innovations include:

1) Design separable multi-scale convolutional blocks and adopt a depth-separable convolutional structure, which significantly reduces the number of parameters and computational load while extracting multi-scale features.

2) Develop broadcast self-attention blocks and replace traditional matrix multiplication with broadcast operations, reducing the computational complexity from quadratic to linear levels.

3) Build a complete lightweight diagnostic framework to significantly enhance computational efficiency while maintaining diagnostic accuracy.

2. Relevant theories

2.1. Multi-scale convolutional module

The structure of the Multiscale Convolution (MC) module is illustrated in Fig. 1 [9]. In contrast to conventional single-scale convolution kernels, the MC block extracts features from multiple local receptive fields by employing convolution kernels of varying sizes in parallel. The resulting features are concatenated, and Batch Normalization (BN) layers are applied to preserve the information distribution across the channel dimension. Ultimately, the GELU (Gaussian Error Linear Unit) activation function is utilized for nonlinear mapping, while retaining a portion of the gradient of negative features. This process can be represented by the following formula:

$$y^{k_l} = Concat_{j=1}^{C_2}\left(\sum_{i=1}^{C_1} w_{i,j}^{k_l} * x_i\right), \tag{1}$$

$$y = GELU(BN(Concat(y^{k_1}, y^{k_2}, \cdots, y^{k_L}))), \tag{2}$$

where, $x \in R^{C_1 \times N_1}$ represents the input; $y \in R^{LC_2 \times N_2}$ indicates the output; $w^{k_l} \in R^{C_2 \times C_1 \times k}$ represents the weight of the lth convolution kernel; $*$ denotes the convolution operation; $Concat(\cdot)$ indicates the splicing operation; C_1 is the dimension of the input channel; N_1 is the input time dimension; C_2 presents the output channel dimension corresponding to each convolution kernel size; N_2 is the output time dimension; k_l is the size of the lth convolution kernel; and L represents the number of different convolution kernel sizes.

Based on the convolution process outlined in Eq. (1), t the learning parameters (Params) and floating-point operations (FLOPs) of the MC block can be obtained, which are expressed as follows:

$$Params_{MC} = \sum_{l=1}^{L} k_l \times C_1 \times C_2, \tag{3}$$

$$FLOPs_{MC} = \sum_{l=1}^{L} k_l \times C_1 \times C_2 \times N_2, \tag{4}$$

where, C_1 represents the number of input channels, C_2 is the number of output channels corresponding to each scale convolution kernel, k_l indicates the size of the lth convolution kernel, L is the number of different convolution kernel sizes, and N_2 is the time dimension of the output sequence.

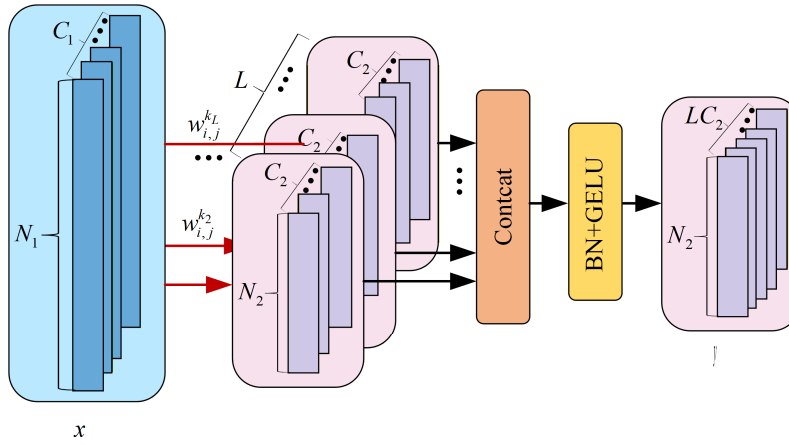


Fig. 1. The block of multiscale convolution

2.2. Multihead self-attention block

The Multihead Self-Attention (MSA) module, a fundamental component of the Transformer [5], is capable of extracting global feature information from the temporal dimension of the data. Its structure is illustrated in Fig. 2. This module initially transforms the input $\chi \in \mathbb{R}^{N\times C}$ into the query matrix q_h, the key matrix κ_h, and the value matrix v_h through three parallel linear transformations. The corresponding expressions are as follows:

$$q_h = \omega_h^q \chi, \quad \kappa_h = \omega_h^\kappa \chi, \quad v_h = \omega_h^v \chi, \tag{5}$$

among them, ω_h^q, ω_h^κ, $\omega_h^v \in \mathbb{R}^{C_H\times C}$ are the weights of the corresponding linear transformations respectively; H represents the number of attention heads, N represents the length of the time dimension, C represents the input feature dimension, C_H represents the feature dimension of each attention head, and it satisfies $C = H \times C_H$.

Subsequently, matrix multiplication is conducted between the query matrix q_h and the key matrix κ_h to derive the global association information. Following scaling and normalization, the attention weight α_h is obtained, expressed as:

$$\alpha_h = Softmax\left(\frac{q_h \kappa_h^T}{\sqrt{C_H}}\right), \tag{6}$$

where, $\alpha_h \in \mathbb{R}^{N\times N}$ represents the attention weight matrix, and T indicates the transposition operation.

Finally, apply the attention weights to the value matrix v_h, and obtain the output through concatenation and linear transformation:

$$\chi' = Concat(\alpha_1 v_1, \alpha_2 v_2, \ldots, \alpha_H v_H)\, \omega^\chi, \tag{7}$$

where, $\chi' \in \mathbb{R}^{N\times C}$ is the output, and $\omega^\chi \in \mathbb{R}^{C\times C}$ is the weight of the linear transformation of the output. Based on the above calculation process, the Params and FLOPs of the MSA module can

be obtained as follows:

$$Params_{MSA} = 4C^2, \tag{8}$$
$$FLOPs_{MSA} = 4NC^2 + 2CN^2 + HN^2, \tag{9}$$

among them, $4NC^2$, $2CN^2$, HN^2 respectively represent the floating-point operations corresponding to linear transformation, matrix multiplication and multi-dimensional exponential operation.

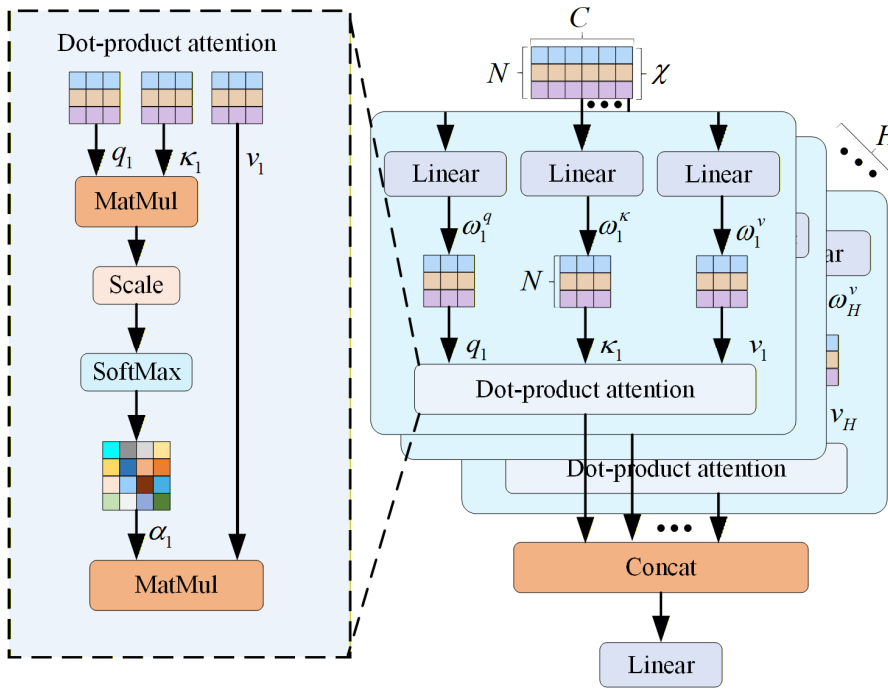


Fig. 2. Multihead self-attention block

3. Proposed method

3.1. Dynamic separable multi-scale convolution

The conventional multi-scale convolution (MC) module is characterized by a large number of parameters and significant computational demands, which hinder its deployment in industrial settings. Furthermore, its rigid architecture presents challenges in adapting to complex fault scenarios. To address these issues, this paper introduces a Dynamic Separable Multi-Scale Convolution (DSMC) block designed to balance model complexity with feature adaptability. The overall structure of the DSMC block is illustrated in Fig. 3 and comprises four primary steps: feature projection, parallel multi-scale depth-separable convolution, dynamic gated weight generation, and weighted fusion.

Firstly, perform a 1×1 convolution operation on the input feature map $x \in \mathbb{R}^{C_{in} \times L}$ to achieve the integration and adjustment of the channel dimensions. This process can be expressed as:

$$Z = Conv1D_{\ k=1}\left(x, W_{proj}\right), \quad Z \in \mathbb{R}^{C_{mid} \times L}, \tag{10}$$

where, W_{proj} denotes the projection weight, and C_{mid} indicates the number of channels in the middle layer. Subsequently, a parallel multi-scale depth-separable convolution structure is

employed. The feature map Z is concurrently input into K paths, each utilizing distinct convolution kernel sizes k_i $(i = 1,2,\ldots,K)$. Each path performs feature extraction through a cascade of depthwise convolution and pointwise convolution. The output i of the Y_i path is computed as follows:

$$Y_i = DepthwiseConv1D_{\,k=k_i}\left(Z, W^i_{depth}\right), \quad Conv1D_{\,k=1}\left(Y_i, W^i_{point}\right), \tag{11}$$

among them, W^i_{depth} and W^i_{point} are the weights of depthwise convolution and pointwise convolution respectively. This decomposition strategy significantly reduces the number of parameters.

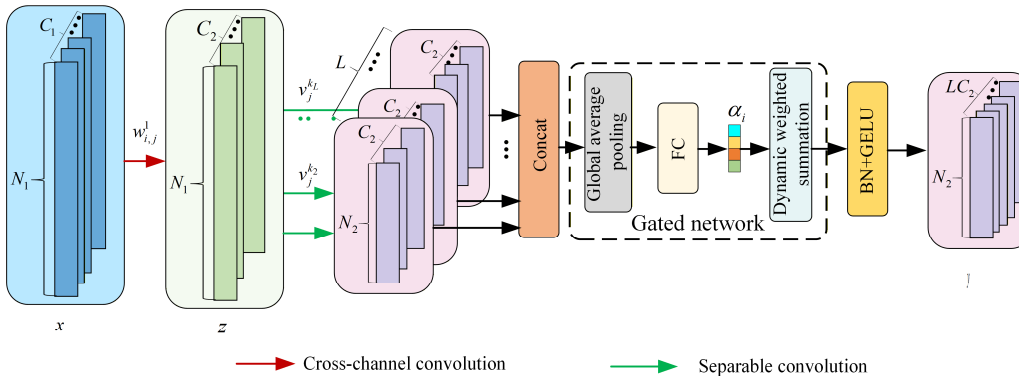


Fig. 3. Dynamic separable multi-scale convolution block

To facilitate adaptive feature fusion, we designed a gated network that incorporates global average pooling and a double-layer fully connected layer. Initially, the network applies global average pooling (GAP) to the input features Z to extract global context information. Subsequently, it generates a set of normalized fusion weights α_i through a micro-network composed of fully connected layers and Softmax functions:

$$\mathbf{g} = GAP(Z), \tag{12}$$
$$\alpha = Softmax(\mathbf{W}_2 \cdot ReLU(\mathbf{W}_1 \cdot \mathbf{g})), \tag{13}$$

among these, $\mathbf{g}$ denotes the channel descriptor obtained by GAP of the input feature map. $\mathbf{W}_1$ and $\mathbf{W}_2$ represent the weights of two 1×1 convolutional layers in the gated network, where r is the channel reduction ratio. K is the number of multi-scale branches, and α_i is the learned dynamic weight for the i-th branch, with the constraint $\sum_{i=1}^{K} \alpha_i = 1$. In the case of simple samples, the gated network effectively suppresses the influence of large-scale convolutional kernels. Conversely, for complex samples, the involvement of all scale convolutional kernels is amplified, particularly through an increase in the weight of large-scale convolutional kernels, facilitating adaptive feature extraction with minimal parameters and negligible increase in model complexity. Ultimately, the output features Y_i from each parallel path are weighted and summed according to the weights α_i produced by the gated network, and the final output Y is obtained via the activation function GELU:

$$Y = GELU\left(\sum_{i=1}^{K} \alpha_i \cdot Y_i\right). \tag{14}$$

This dynamic fusion mechanism enables the model to autonomously emphasize or suppress the contribution of specific scale features based on the characteristics of the input samples.

According to Eq. (10-13), the Params and FLOPs of the DSMC block can be obtained:

$$Params_{DSMC} \approx C_1 \times C_2 + \sum_{l=1}^{L} k_l \times C_2,$$
$$FLOPs_{DSMC} \approx C_1 \times C_2 \times N_1 + \sum_{l=1}^{L} k_l \times C_2 \times N_2, \tag{15}$$

where, C_1 and C_2 denote the input and output channel numbers of the DSMC module, respectively. L is the number of parallel multi-scale convolution branches, k_l is the size of the l-th convolution kernel, and N_1, N_2 represent the temporal sequence lengths of input and output feature maps, respectively. It can be seen from Eqs. (3-4) that the degree to which the DSMC block reduces Params compared to the ordinary MC block is respectively related to the values of $\sum_{l=1}^{L} k_l$ and C_1. Similarly, the degree of FLOPs is also related to $\sum_{l=1}^{L} k_l \times \frac{N_2}{N_1}$ and the value of C_1.

3.2. Broadcast self-attention

The conventional multi-head self-attention (MSA) module depends on intricate matrix multiplications and multi-dimensional exponential operations, which complicate its ability to satisfy lightweight requirements. Drawing inspiration from the separable Self-Attention mechanism [11], this paper introduces the Broadcast Self-Attention (BSA) module, as illustrated in Fig. 4. This module requires traversal of only one-dimensional time series data. By incorporating broadcast operations to facilitate global information transmission, the computational complexity is markedly diminished.

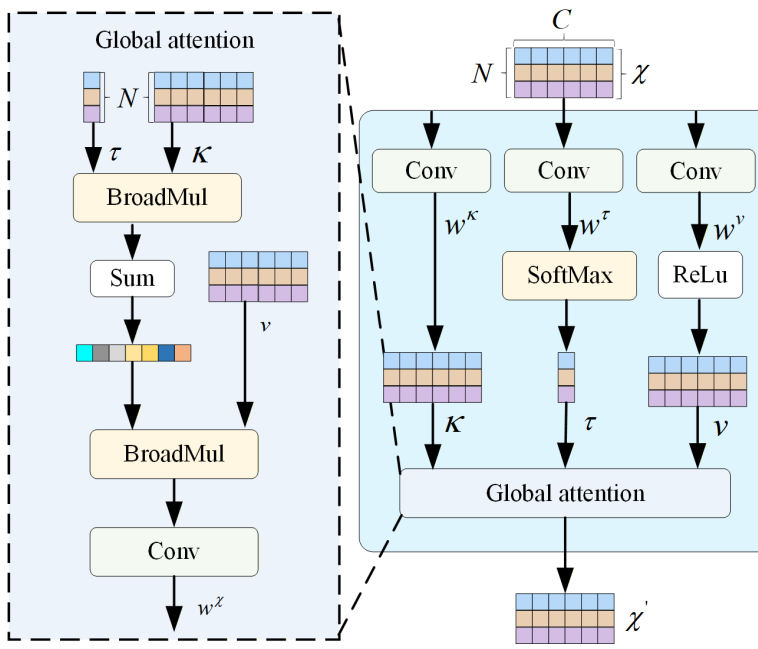


Fig. 4. Broadcast self-attention

First, calculate the time fraction τ, key matrix κ and value matrix v through a convolution operation with a kernel size of 1:

$$\tau = \text{Softmax}(w^{\tau} * \chi), \quad \kappa = w^{\kappa} * \chi, \quad v = ReLU(w^{v} * \chi), \tag{16}$$

among them, w^{τ}, w^{κ}, w^{v} is the convolution weight, and κ, $v \in \mathbb{R}^{N\times C}$ is respectively used for feature mapping and nonlinear transformation. Subsequently, the time score τ is assigned to each dimension of the key matrix κ through the broadcasting mechanism, and the fine-grained feature weights γ are obtained by summing in the time dimension:

$$\gamma = \sum_{i=1}^{N} (\tau_i \otimes \kappa_i), \tag{17}$$

where, $\otimes$ represents the broadcast operation. This weight is again broadcast to the value matrix v to extract the global key features. The final output is linearly mapped through 1×1 convolution:

$$\chi' = \sum_{j=1}^{C} (\gamma_j \otimes v_j) * w^{\chi}, \tag{18}$$

where, $w^{\chi} \in \mathbb{R}^{C\times C}$ represents the convolution weight. According to the above calculation process, the Params and FLOPs of the BSA module are respectively:

$$\begin{aligned} Params_{BSA} &= 3C^2 + C, \\ FLOPs_{BSA} &= N(3C^2 + C) + 2CN + N, \end{aligned} \tag{19}$$

among them, the complexity of linear transformation, broadcast operation and single-dimensional exponential operation has all been reduced to $O(N)$, which is lower than that of the MSA module, significantly improving the computational efficiency.

In the standard self-attention mechanism, the pairwise similarity calculation between query and key vectors relies on dense matrix multiplication QK^T, which inevitably brings inherent $O(N^2d)$ computational complexity with the increase of sequence length N. To address this issue, this paper reconstructs the attention calculation process based on broadcast operation, and achieves the reduction of computational complexity from $O(N^2)$ to $O(N)$.

In terms of mathematical equivalence and information capacity, the broadcast-based attention operation proposed in this paper is strictly equivalent to traditional matrix multiplication in self-attention. The broadcast mechanism only decomposes the dense matrix multiplication into element-wise vector operations along the dimension direction, which does not change the mapping space of attention similarity weight, nor reduces the information capacity of original query, key and value features. The feature entropy and feature distribution coverage of the attention output are completely consistent with the standard self-attention. In other words, the broadcast reconstruction eliminates redundant square-level computation caused by global pairwise dot products, while fully retaining the original feature expression ability and information transmission capacity of the attention mechanism.

The complexity optimization principle is further analyzed as follows. The dense matrix multiplication of standard self-attention generates an $N \times N$ global similarity matrix, which is the source of $O(N^2)$ complexity. By virtue of broadcast expansion and element-wise operation, the proposed BSA module avoids the construction of dense similarity matrix, and all calculations only perform linear traversal along the sequence dimension. Therefore, the overall computational complexity is reduced to linear $O(Nd)$, realizing efficient lightweight attention feature extraction without information loss.

3.3. Lightweight LWConvFormer fault diagnosis framework

As illustrated in Fig. 5, LWConvFormer employs a hierarchical encoder architecture comprising three components: the signal preprocessing module, the feature extraction backbone

network, and the classification output module. This framework, designed with systematic lightweight principles, markedly decreases computational complexity while maintaining diagnostic accuracy, making it well-suited for real-time monitoring applications in industrial environments.

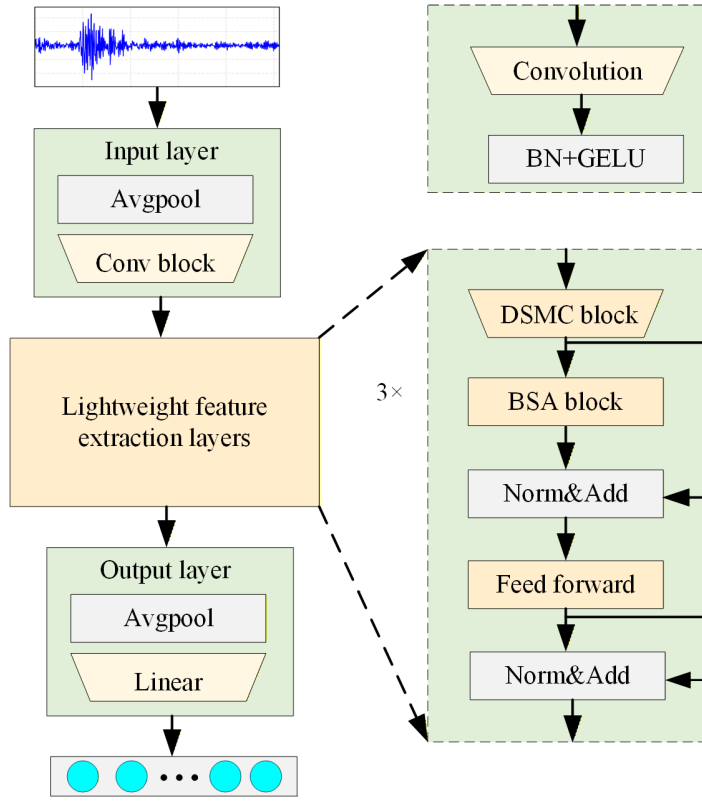


Fig. 5. Structure of LWConvFormer

During the preprocessing stage, the raw vibration signal is first subjected to temporal down-sampling via an average pooling layer to mitigate high-frequency noise and reduce computational overhead. Subsequently, a one-dimensional convolutional layer is employed to facilitate local feature fusion and project the signal into a high-dimensional latent space. This initial processing not only retains the essential time-domain characteristics but also establishes a robust foundation for subsequent deep feature extraction T The feature extraction backbone adopts a three-layer progressive encoder architecture. Each layer hierarchically integrates dynamic separable multi-scale convolution (DSMC) to capture multi-granularity fault patterns, followed by linear-complexity broadcast self-attention (BSA) to model global contextual dependencies. These operations are stabilized by layer normalization and residual connections, and further refined by a lightweight feedforward network [12]. This design effectively balances local perceptual sensitivity with global correlation modeling. Finally, the classification head utilizes global average pooling to compress spatial redundancy and suppress overfitting, ultimately mapping the learned representations to fault categories via a linear classifier.

Fig. 6 illustrates the diagnostic process of this framework, which comprises three steps. Initially, vibration signals are gathered, and samples are partitioned into training, validation, and test sets using sliding windows. Next, lightweight training is performed utilizing cross-entropy loss and the AdamW optimizer [13], and the best model is chosen based on the validation set's performance. Lastly, the diagnostic accuracy is assessed on the test set, and the model's

generalization capability is thoroughly examined in conjunction with visualization techniques.

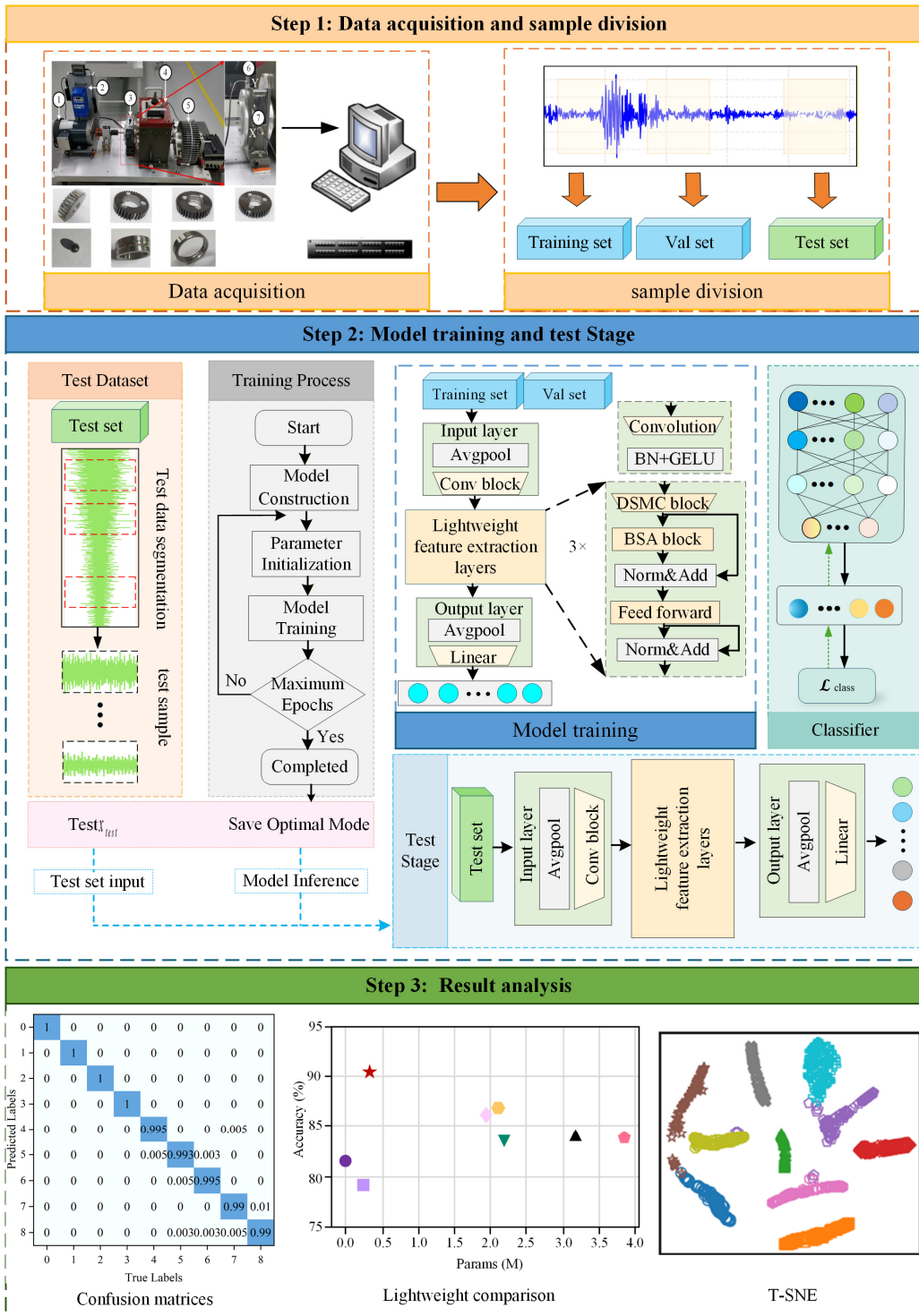


Fig. 6. Lightweight fault diagnosis framework

ISSN PRINT 1392-8716, ISSN ONLINE 2538-8460

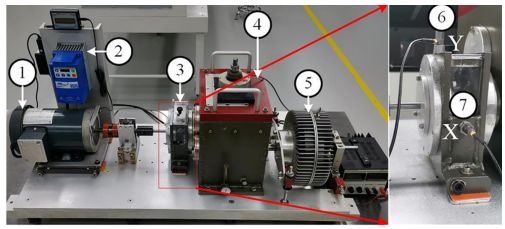
4. Experimental section

This section assesses the validity of the model using the planetary gearbox fault diagnosis experimental platform at Xi'an Jiaotong University and the QPZZ-II type rotating machinery test bench at East China Jiaotong University. The experiment was conducted on the PyTorch deep learning framework and accelerated using CUDA 11.8. The hardware configuration included an Intel64 Family 6 Model 140 Stepping 1 CPU and an NVIDIA GeForce MX450 graphics card.

4.1. Case 1: Planetary gearbox fault diagnosis case

4.1.1. Dataset description

This study employs the planetary gearbox dataset from Xi'an Jiaotong University [11], with two vibration channels collected by ICP acceleration sensors in the X and Y directions. The experimental setup, as shown in Fig. 7, operates at a rotating speed of 1800 r/min and a sampling frequency of 20,480 Hz. The dataset includes 9 health conditions consisting of 1 normal state, 4 typical gear faults, and 4 bearing faults. For sample construction, a fixed-length segmentation with non-overlapping intervals is applied. Each sample is generated using a segment length of 1024 sampling points for each channel, and the step size between consecutive samples is set to 1500 sampling points. The window length of 1024 points is selected because it covers more than one complete shaft rotation cycle under the working condition, which can fully retain the periodic fault characteristics in the vibration signal. Two-channel signals are concatenated along the channel dimension to form a 2×1024 sample. As shown in Table 1, a total of 1200 valid samples are generated for each health condition, leading to 10,800 samples in total. The dataset is divided into training, validation, and test sets with a stratified split ratio of 5:3:4 to ensure consistent class distribution across subsets.



1. Motor, 2. Controller, 3. Planetary gearbox, 4. Parallel gearbox, 5. Brake, 6-7. Accelerometor in horizontal, and vertical direction.

a)



b)

Fig. 7. Gearbox test bench: a) test bench; b) categories of gear and bearing faults. Photo by Xi'an Jiaotong University, Xi'an, Shanxi, China, 2025

Table 1. The diagnostic accuracy and model complexity of each model in Case 1

Fault Category	Fault Size / mm	Number of samples	Label
Normal	/	1200	0
Slight gear wear	0.2	1200	1
Medium gear wear	0.4	1200	2
Severe gear wear	0.6	1200	3
Gear broken tooth	Complete single-tooth breakage	1200	4
Slight bearing inner race fault	0.3	1200	5
Medium bearing inner race fault	0.6	1200	6
Severe bearing inner race fault	0.9	1200	7
Bearing outer race fault	0.6	1200	8

To simulate an industrial noise environment, two types of noise were introduced to the test set

to assess the method's robustness against noise interference. Random Gaussian noise:

$$s_i = s_i + \varepsilon. \tag{20}$$

Random scaling noise:

$$s_i = \sigma \times s_i, \tag{21}$$

where, $\varepsilon \sim N(0, \lambda)$, $\sigma \sim N(1, \lambda)$, s_i represent the ith sampling point of the sample, and λ is the noise variance. Each type of noise is randomly added with a 50 % probability, and different intensities of noise interference are simulated by adjusting the values.

4.1.2. Result analysis

To systematically assess the effectiveness of the proposed method, this study selected three representative comparison approaches: 1) new end-to-end fault diagnosis methods utilizing the CNN-Transformer fusion architecture (CLFormer [9], Conformer-NSE [14], MCSwin-T [10]); and 2) four classic CNN architectures (MobileNet [15], MobileNet-V2 [16], ResNet18 [17], MK-ResCNN [18]). All experiments were conducted with five repetitions to mitigate the effects of randomness. The training duration was set to 100 epochs, and the batch size was consistently maintained at 32.

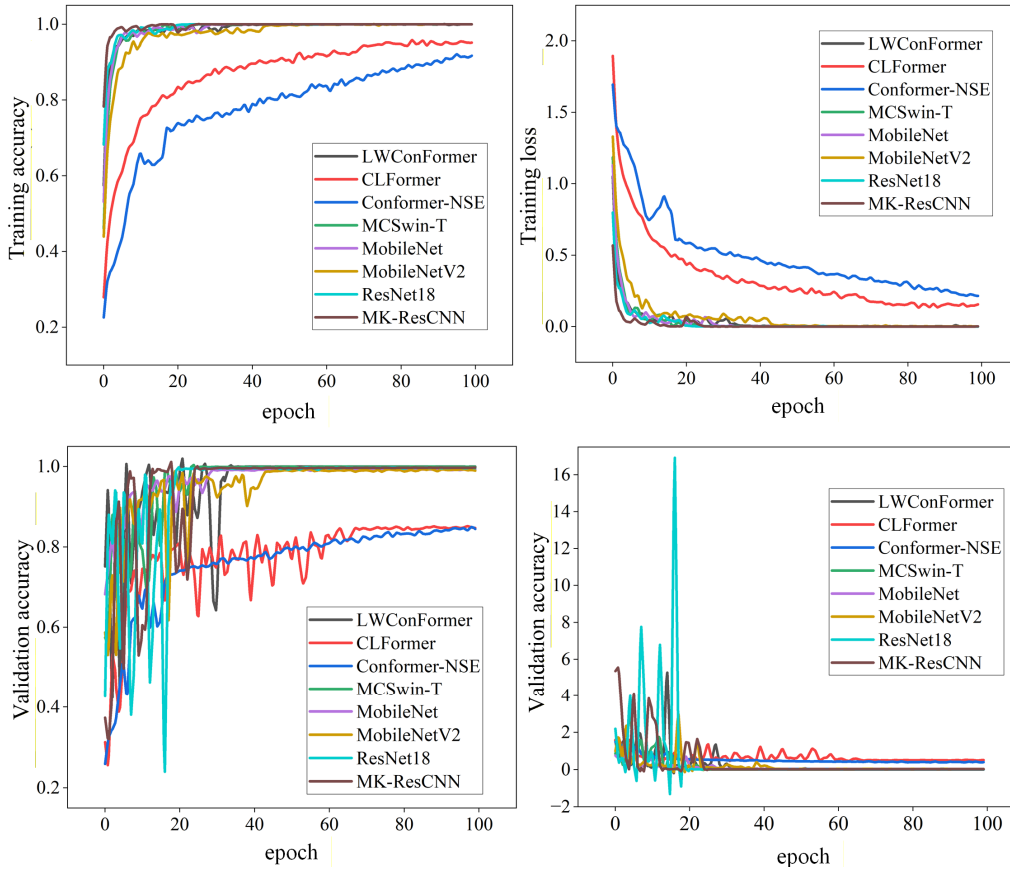


Fig. 8. The accuracy and loss curves of training and validation

Fig. 8 shows the training process curves of each method. It can be seen that in the early stage of training, the fluctuations of all methods on the training set are milder than those on the validation set. After approximately 80 rounds of iterations, all methods tend to converge on both the training set and the validation set, presenting stable low losses and high accuracy. It is worth noting that the accuracy of the validation set after convergence of CLFormer and Conformer-NSE is significantly low. This might be due to the feature dimension compression strategy they adopt, which leads to the loss of fine-grained features in the signal, thereby weakening the complete learning ability of the model for multi-dimensional information.

Table 2 shows the performance of each diagnostic method at different noise levels. Some comparison models perform slightly better under noise-free conditions, but their anti-noise performance is relatively poor. With the increase of noise intensity, the anti-noise advantage of the method proposed in this paper gradually becomes obvious, and its diagnostic performance remains leading at both noise levels. Specifically, when $\lambda = 0.2$, the average and highest accuracy rates of the suboptimal method are 0.86 % and 1.11 % lower than those of the method proposed in this paper, respectively. When $\lambda = 0.4$, the gap expands to 4.56 % and 4.42 %. In terms of model efficiency, although CLFormer and Conformer-NSE have lightweight characteristics, their feature compression processes sacrifice fine-grained features, leading to a notable trade-off between complexity and performance. Taking CLFormer as an example, while it achieves the smallest parameter count (only 0.005 M), its diagnostic accuracy drops sharply under challenging conditions, reaching only 77.16 % at $\lambda = 0.4$. This indicates that an overly aggressive reduction in model complexity inevitably weakens the model's feature representation capability, resulting in a significant performance degradation in noisy and variable working environments. In contrast, the method proposed in this paper achieves a good balance between model complexity and operational efficiency. It significantly reduces computational overhead while maintaining high diagnostic performance, effectively avoiding the "over-lightweighting trap" that plagues models such as CLFormer. Experiments show that the proposed framework, on the basis of lightweight design, has strong anti-noise and feature extraction capabilities, providing a feasible solution for fault diagnosis in industrial sites.

Table 2. The diagnostic accuracy and model complexity of each model in Case 1

Model	Accuracy (%)						Complexity (M)		Time (s)
	$\lambda = 0$		$\lambda = 0.2$		$\lambda = 0.4$				
	Mean	Max	Mean	Max	Mean	Max	Param	FLOPs	Training
LWConvFormer	99.67	100	96.88	98.25	85.93	91.42	0.323	14.520	259.17
CLFormer	92.70	98.53	87.43	92.06	77.16	84.72	0.005	0.144	272.84
Conformer-NSE	94.07	98.14	85.03	88.92	73.04	75.11	0.245	6.270	342.47
MCSwin-T	99.91	99.94	95.52	97.78	80.52	84.50	1.937	227.003	1058.2
MobileNet	99.67	99.83	91.80	93.39	76.44	78.92	3.186	333.620	1412.3
MobileNetV2	99.66	99.81	90.82	91.89	75.34	76.25	2.192	96.955	815.55
ResNet18	99.66	99.78	94.67	95.00	78.76	80.50	3.854	175.920	1090.8
MK-ResCNN	99.70	99.83	96.02	97.14	81.37	87.00	2.117	83.893	450.64

As illustrated in Fig. 9, this paper further conducts a detailed comparative analysis of the predicted labels and true labels of all comparative models based on normalized confusion matrices, identifies the fault categories that are prone to misclassification in the experiment, and explains the causes of misclassification by combining the physical mechanism of faulty vibration signals. As can be observed from the confusion matrix results, all models achieve nearly perfect recognition of normal operating conditions and gear wear faults, with almost no inter-class misclassification. The misclassification issues of all models are concentrated between bearing inner race faults and bearing outer race faults; among these, severe bearing inner race faults and outer race faults are extremely prone to mutual confusion, which constitutes the primary misclassification condition in this gearbox fault diagnosis task. From the perspective of physical mechanism, under the same operating conditions, the vibration signal characteristics excited by

bearing inner and outer race damages are highly similar. The time-frequency harmonic components generated by the two types of fault excitations overlap significantly, resulting in ambiguous boundaries of fault features, which in turn leads to category confusion in the model. Although the overall classification accuracy of the proposed model is not the highest among all comparative methods, it maintains a low misclassification ratio for high-similarity fault pairs and features the lowest model architecture complexity. The results demonstrate that the proposed lightweight diagnostic framework possesses stable robustness in feature extraction and reliable engineering generalization performance, rather than unilaterally pursuing a single accuracy metric.

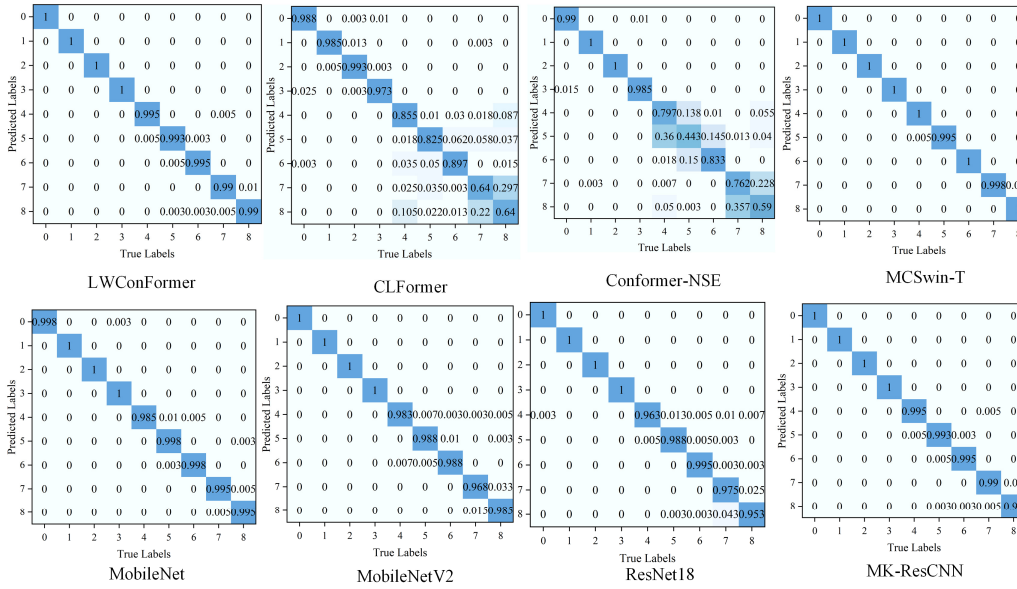


Fig. 9. Confusion matrices of all models under noise intensity $\lambda = 0$

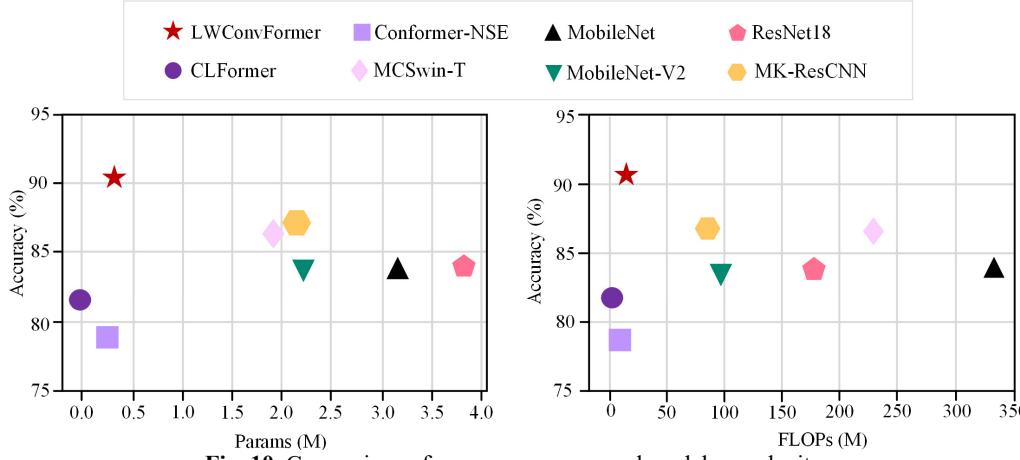


Fig. 10. Comparison of average accuracy and model complexity

Fig. 10 provides a visual comparison of the diagnostic performance and model complexity of each method across varying noise levels. The findings indicate that the diagnostic performance and robustness of the method proposed in this paper surpass those of existing approaches while preserving model lightweightness. Specifically, under two noise conditions, the proposed method

ISSN PRINT 1392-8716, ISSN ONLINE 2538-8460

(with a parameter count of 0.323 M and a computational load of 14.520 M FLOPs) achieved an average accuracy rate of 94.16 %. In comparison to the suboptimal MCSwin-T, the proposed method enhances the accuracy rate by 1.47 %, while reducing the number of parameters and computational load to one-sixth and less than one-tenth, respectively. When compared to the more effective CNN method MK-ResCNN, the proposed method increases the accuracy rate by 1.8 %, while also reducing both the number of parameters and the computational load to approximately one-sixth of MK-ResCNN. This comprehensive comparison clearly illustrates the advantages of the proposed method regarding model efficiency and diagnostic performance.

For the practical demand of real-time fault diagnosis in industrial scenarios, the inference efficiency of the model is the core indicator to measure its deployment feasibility. To verify the real-time performance of the proposed model, this paper supplements the test of single-sample inference latency and frames per second (FPS) of each comparison model. All experiments are conducted in the same hardware environment, and the batch size is set to 1 to fit the real scenario of online diagnosis. The test results are shown in Table 3.

It can be seen from Table 3 that there are significant differences in the inference performance of each model. Among them, ResNet18 and MobileNet achieve low inference latency (2.464 ms and 2.654 ms, respectively) and high processing efficiency (405.8 FPS and 376.8 FPS, respectively) relying on mature network structure design, but they are difficult to adapt to the deployment scenarios with limited resources in industrial fields. The inference latency of CLFormer reaches 4.289 ms, and the processing speed is relatively slow, which is difficult to meet the demand of high-frequency real-time monitoring. The inference latency of mcswint, MobileNetV2, and MSResNet is between 3.214 ms and 5.914 ms, with relatively limited processing efficiency, which is difficult to adapt to the demand of high real-time diagnosis.The proposed LWconvformer model shows excellent inference performance: its single-sample inference latency is 3.590 ms, and it can process 278.5 samples per second. Although it is slightly higher than ResNet18 and MobileNet, it fully meets the real-time requirements of online fault diagnosis in industrial fields (usually requiring inference latency less than 10 ms). In addition, combined with the diagnostic accuracy experimental results mentioned earlier, LWconvformer can still maintain excellent fault diagnosis performance while ensuring high inference efficiency, effectively solving the pain point that traditional models “are difficult to balance accuracy and efficiency”.

Table 3. Test results of real-time inference performance of comparison models in Case 1

Model	Inference latency (ms)	FPS
LWConvFormer	3.590	278.5
CLFormer	4.289	233.1
Conformer-NSE	5.877	170.2
MCSwin-T	5.914	169.1
MobileNet	2.654	376.8
MobileNetV2	5.187	192.4
ResNet18	2.464	405.8
MK-ResCNN	3.214	311.2

In conclusion, LWconvformer achieves an optimal balance between inference latency, processing efficiency, and diagnostic accuracy. It not only has sufficient inference efficiency to meet real-time deployment needs but also can maintain excellent diagnostic performance, which fully verifies its practicality and superiority in the field of industrial equipment fault diagnosis.

To verify the adaptive feature extraction capability of the proposed dynamic gating mechanism, we visualize the learned gate weight distributions across different fault classes, as shown in the Fig. 11. Class 0 corresponds to the normal condition, Classes 1-3 correspond to spur gear faults, Classes 4-6 correspond to bearing faults, and Classes 7-8 correspond to planetary gear faults. The results reveal that the model exhibits significantly different scale preferences for different fault classes, which is consistent with the adaptive weighting behavior of the gated

network described in Section 3.

Under the normal condition, the model assigns the highest proportion of weight to medium-scale features (Scale 5). This indicates that for stationary vibration signals, the gated network – generating weights via global average pooling (GAP) and a two-layer fully connected structure – adaptively increases the importance of medium-scale features to capture subtle fluctuations, thereby distinguishing normal operating conditions from incipient faults. For spur gear faults, the weight proportion of large-scale features (Scale 7) is significantly higher, which is highly consistent with the low-frequency periodic impact characteristics of gear meshing faults. This demonstrates that the gating mechanism can automatically focus on large-scale features that reflect the meshing frequency and its harmonics, by learning to assign higher weights to relevant branches through Softmax normalization. For bearing faults, the weight proportion of small-scale features (Scale 3) is significantly increased. This shows that the gated network pays more attention to the high-frequency impact details generated by bearing defects, assigning larger weights to small-scale branches to enhance the discriminability of these fault patterns. For planetary gear faults, the model assigns the highest weight proportion to medium-scale features (Scale 5), reflecting its adaptability to the complex motion patterns of planetary gears, which involve both low-frequency rotation and high-frequency meshing components.

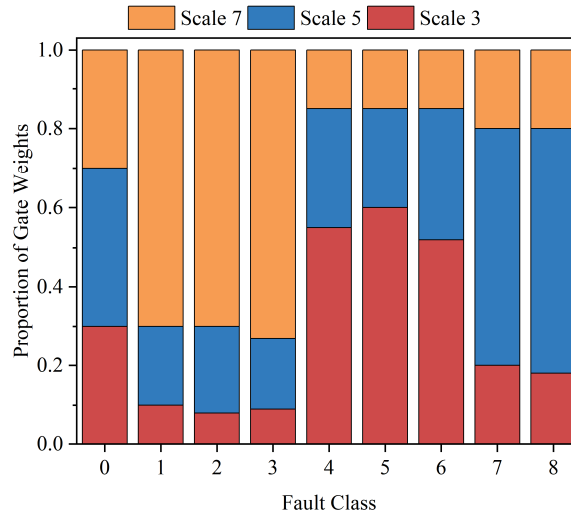


Fig. 11. Gate weight distributions across different fault classes

This dynamically adjusted weight distribution pattern across fault classes fully verifies that the proposed dynamic gating mechanism, which generates normalized fusion weights via GAP, ReLU-activated fully connected layers, and Softmax, can adaptively adjust the importance of multi-scale features according to the fault characteristics of input signals, thus achieving the design purpose of adaptive feature extraction.

4.1.3. Ablation study analysis

To quantitatively analyze the individual contributions of the DSMC multi-scale convolution module and BSA broadcast self-attention module to diagnostic performance and lightweight characteristics of the proposed framework, three groups of controlled ablation experiments are carried out according to the module structure. The results of average diagnostic accuracy, model parameters and floating-point computation of each structure are shown in Table 4.

In the first group (DSMC + standard MSA), the original DSMC module is retained, while the BSA module is replaced by traditional O(N2) standard self-attention. This structure achieves an

average accuracy of 98.03 % with 0.409 M parameters and extremely high FLOPs of 18.165 M. The results demonstrate that traditional dense matrix-multiplication attention brings enormous computational overhead and obvious accuracy degradation, which reveals the information redundancy and performance loss of standard MSA in long-sequence fault signal modeling.

Table 4. The diagnostic accuracy and model complexity of each model in case 1

Model structure	Ablation mode	Average accuracy (%)	Parameters (M)	FLOPs (M)
DSMC + standard MSA	Ablation_a	98.03	0.409	18.165
Standard MC + BSA	Ablation_b	99.47	0.955	40.980
Full LWConvFormer	Full	99.67	0.323	14.520

In the second group (standard MC + BSA), the BSA broadcast attention is maintained, and the DSMC module is replaced with vanilla single-scale MC convolution. The accuracy of this structure is improved to 99.47 %, which verifies the powerful feature modeling capability of the BSA module. Nevertheless, its parameters surge to 0.955 M and FLOPs rise sharply to 40.980 M. It proves that vanilla single-scale convolution fails to extract multi-scale fault features, and leads to redundant model structure and explosive computational complexity, which highlights the necessity of lightweight multi-scale feature extraction of the DSMC module.

In the third group, the complete proposed LWConvFormer framework (DSMC + BSA) is adopted. The integrated model achieves the highest average diagnostic accuracy of 99.67 % among all ablation groups, with the minimum parameter quantity of 0.323 M and low FLOPs of 14.520 M.

In conclusion, the DSMC module effectively compresses redundant model parameters and captures multi-scale vibration features of fault signals. The BSA module reconstructs attention computation with linear complexity, which eliminates the quadratic computational cost of traditional attention while fully preserving feature information modeling ability. The synergistic combination of the two modules realizes the optimal trade-off among diagnostic accuracy, parameter scale and computational complexity, which fully verifies the rationality of module design and the superior lightweight diagnosis performance of the proposed framework.

4.2. Case 2: QPZZ-II type rotating machinery test bench

4.2.1. Dataset description

The dataset utilized in this study is sourced from the QPZZ-II type rotating machinery test bench at East China Jiaotong University. The configuration of this test bench is illustrated in Fig. 12 and comprises a drive motor, mechanical transmission, and loading and fixing devices. By adjusting the installation positions of components and their combinations, the test bench can effectively simulate various faults in rotating machinery within a short timeframe. During data acquisition, the sampling frequency is set to 12 kHz, and the constant operating speed of the bearing is 1000 r/min under steady load condition. The test subjects consist of single row cylindrical roller bearings, specifically models N205EM and NU205EM. The detailed parameters of the tested bearings are presented in Table 5.

To ensure the dataset's validity and rationality, a non-overlapping segmentation strategy with a fixed stride of 1024 sampling points is adopted to extract samples from diverse vibration signals. The window length of 1024 points is selected because it fully covers multiple complete shaft rotation cycles under 400 r/min operating condition, which guarantees that periodic fault impulse features in the vibration signal are sufficiently captured. Based on the collected data, the experimental sample set described in Table 6 is established, which covers 10 different bearing health states with an equal number of samples for each category. Each sample comprises 1024 sampling points, and the dataset is partitioned into training, validation, and test sets in a ratio of 7:1:2.

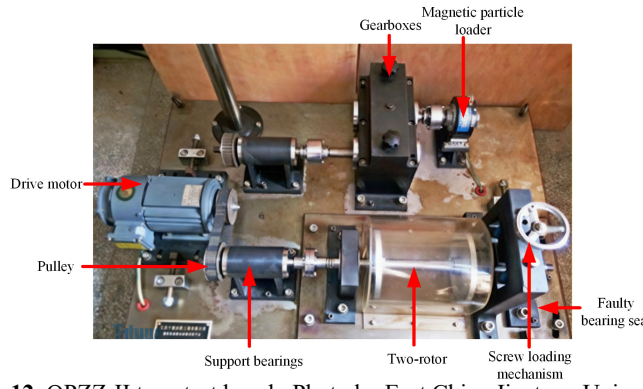


Fig. 12. QPZZ-II type test bench. Photo by East China Jiaotong University, Nanchang, Jiangxi, China, 2025

Table 5. The specific parameters of the tested bearing

Parameters	Value
Pitch circle diameter	38.5 mm
Rolling element diameter	6.75 mm
Outer ring outer diameter	52 mm
Inner ring inner diameter	25 mm
Number of rolling elements	13
Contact angle	0°
Width	15 mm
Limiting rotational speed	14000 rpm

Table 6. Information of experimental samples for the tested bearing

Fault category	Fault size / mm	Number of samples	Label
Normal condition	–	214	C0
Rolling element fault (slight)	0.5	214	C1
Rolling element fault (medium)	1	214	C2
Rolling element fault (severe)	2	214	C3
Inner race fault (slight)	0.5	214	C4
Inner race fault (medium)	1	214	C5
Inner race fault (severe)	2	214	C6
Outer race fault (slight)	0.5	214	C7
Outer race fault (medium)	1	214	C8
Outer race fault (severe)	2	214	C9

4.2.2. Result analysis

Seven comparison methods from Case 1 were selected and evaluated against the method proposed in this paper for performance assessment. The experimental settings mirrored those utilized in Case 1. The results are presented in Table 7 and Fig. 13. The LWConvFormer method introduced in this paper exhibits significant advantages in both noise robustness and model efficiency.

As the noise level increased from 0 to 0.4, the average accuracy of each method generally declined; however, the proposed method maintained the highest accuracy, achieving 94.01 %, 90.06 %, and 84.56 %, respectively. Notably, under conditions of strong noise (λ = 0.4), its accuracy markedly surpassed that of the suboptimal method MCSwin-T. In contrast, methods such as CLFormer can attain a peak accuracy of 95.23 % in the absence of noise, but their average accuracy fluctuates considerably, indicating relatively weak noise robustness. Moreover, despite its extremely small parameter count of only 0.005 M, CLFormer's diagnostic accuracy drops sharply under challenging conditions, reaching merely 64.02 % at λ = 0.4. This highlights a critical

limitation: overly aggressive lightweighting sacrifices feature representation capability, leading to a severe trade-off between complexity and performance. Regarding model efficiency, the parameter count of the proposed method is only 0.322 M, with a computational cost of 14.397 M FLOPs, significantly lower than that of traditional CNN architectures, such as the MobileNet series (2.193-3.187 M), and the Transformer architecture MCSwin-T (1.936 M). Furthermore, the training time is merely 2,124 seconds, nearly seven times faster than that of MobileNetV2, thereby underscoring the superiority of its lightweight design. Unlike CLFormer, which falls into the "over-lightweighting trap", LWConvFormer strikes an optimal balance between model size, computational efficiency, and diagnostic performance, making it more practical for real-world industrial deployment.

Table 7. The diagnostic accuracy and model complexity of each model in Case 2

Model	Accuracy (%)						Complexity (M)		Time (s)
	$\lambda = 0$		$\lambda = 0.2$		$\lambda = 0.4$				
	Mean	Max	Mean	Max	Mean	Max	Param	FLOPs	Training
LWConvFormer	94.01	96.21	90.06	93.12	84.56	90.02	0.322	14.397	2124
CLFormer	68.47	95.23	67.62	68.02	64.02	65.61	0.005	0.154	3065
Conformer-NSE	80.73	82.63	78.91	79.99	72.69	74.01	0.239	6.221	3700
MCSwin-T	94.00	95.96	89.02	90.21	87.01	89.20	1.936	226.507	9707.3
MobileNet	91.54	92.12	86.36	89.35	83.78	84.54	3.187	333.523	12568
MobileNetV2	89.65	91.27	84.23	85.45	81.23	83.65	2.193	96.907	16171
ResNet18	91.27	92.30	86.02	88.21	84.58	85.69	3.854	175.691	14202
MK-ResCNN	91.45	92.61	85.42	87.02	83.21	84.78	2.118	83.664	4332.4

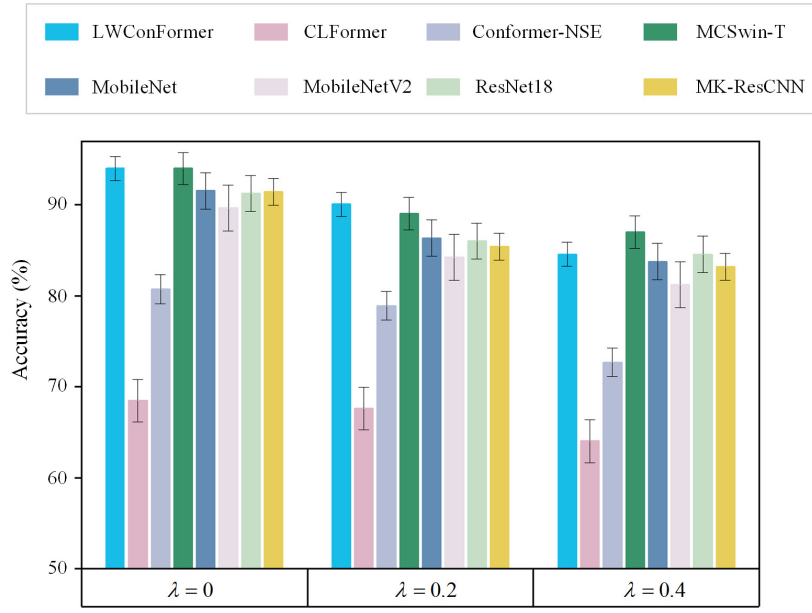


Fig. 13. The accuracy rate under different noises in Case 2

Fig. 14 illustrates the visualization results derived from T-SNE features, which comprehensively reflect the diagnostic performance of each model on the QPZZ-II test bench in the feature space. As shown in the figure and verified quantitatively in Table 8, the proposed LWConvFormer exhibits superior feature learning capabilities compared with other methods. Regarding feature distribution, LWConvFormer achieves the optimal balance between intra-class aggregation and inter-class separation. The feature points of each fault category form compact clusters with clear boundaries. Notably, critical fault categories such as C1, C3, and C5 are densely

packed in the feature space, showing strong separability from other classes. This pattern is further confirmed by the quantitative metrics: LWConvFormer achieves the lowest intra-class distance (0.121) and the highest inter-class distance (2.451), demonstrating its strong ability to capture discriminative fault features even under noisy conditions. In contrast, traditional CNN models such as MobileNetV2 and ResNet18 show satisfactory performance in some categories but suffer from significant feature overlap, reflected in their higher intra-class distances and lower inter-class separability. Transformer-based models, including CLFormer and Conformer-NSE, exhibit indistinct category boundaries and dispersed feature distributions, with CLFormer showing the highest intra-class distance (0.382) and lowest inter-class distance (1.634). These deficiencies in feature representation directly lead to performance degradation in real-world diagnostic tasks.

Table 8. Intra-class and inter-class distance metrics of feature representations for different models

Model	Intra-class distance	Inter-class distance
LWConvFormer	0.121	2.451
CLFormer	0.382	1.634
Conformer-NSE	0.297	1.872
MCSwin-T	0.215	2.013
MobileNet	0.273	1.765
MobileNetV2	0.311	1.598
ResNet18	0.248	1.904
MK-ResCNN	0.269	1.827

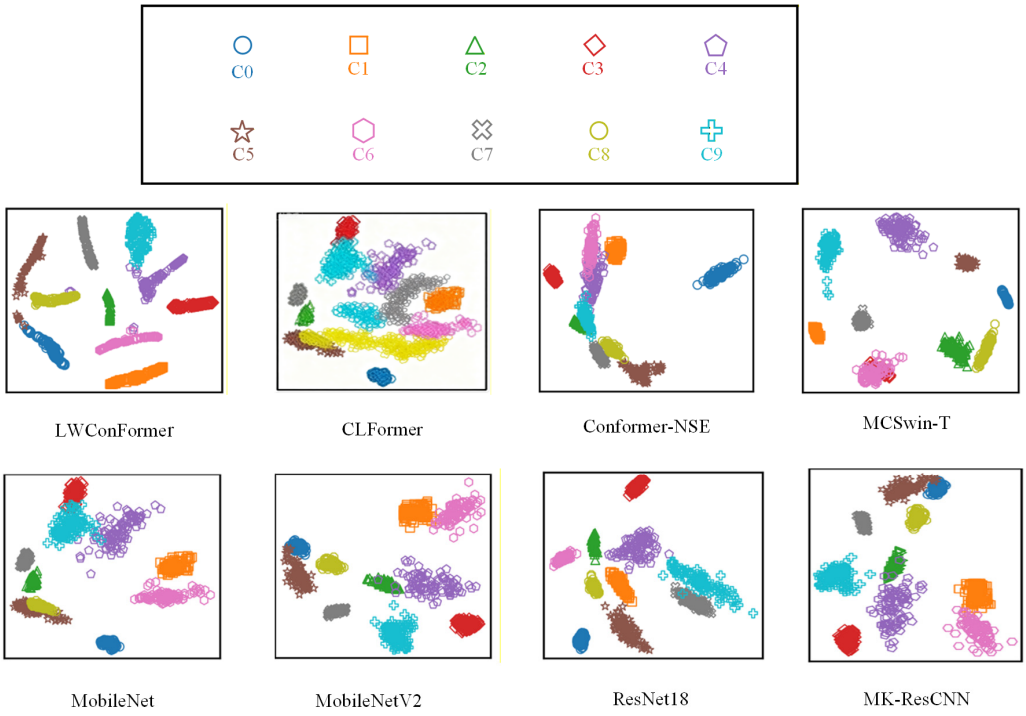
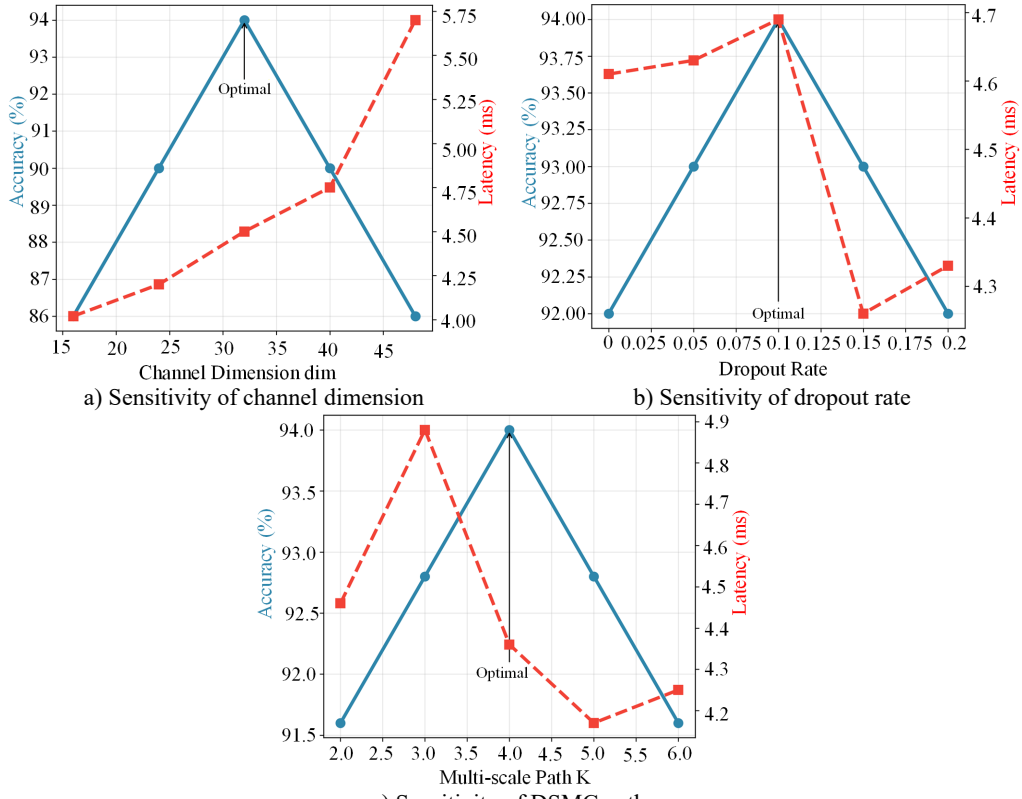


Fig. 14. T-SNE feature visualization diagram

To investigate the influence of key hyperparameters on the performance of the proposed LWConvFormer and to determine their optimal configurations, a comprehensive sensitivity analysis was conducted on the channel dimension dim, the dropout rate, and the number of multi-scale paths K in the DSMC module. The results, as illustrated in Fig. 15, reveal a consistent "rise-then-fall" trend in diagnostic accuracy with variations in these hyperparameters. For the channel dimension, the model's feature representation capability is insufficient at low values,

ISSN PRINT 1392-8716, ISSN ONLINE 2538-8460

leading to lower accuracy. As the dimension increases, accuracy improves significantly; however, excessive dimensions introduce redundant parameters, causing a gradual performance decline. The optimal value is achieved at dim = 32. The dropout rate, which primarily affects generalization ability, yields the highest accuracy at 0.1. A lower rate fails to mitigate overfitting, while a higher rate leads to the loss of critical features. For the multi-scale path number K, a value of 4 balances the sufficiency of multi-scale feature extraction and computational efficiency, avoiding both under-extraction and redundant feature interference observed at lower and higher values, respectively. Across all three hyperparameters, the optimal configurations converge at the peak accuracy points, while the corresponding inference latency remains within a reasonable range. This confirms that the selected hyperparameter combination strikes the best balance between diagnostic accuracy, generalization capability, and computational efficiency, validating the rationality of the design choices.



a) Sensitivity of channel dimension

b) Sensitivity of dropout rate

c) Sensitivity of DSMC path

Fig. 15. Hyperparameter sensitivity analysis of the proposed LWConvFormer

To further verify the generalization and real-time deployment capability of the proposed model, we supplement the inference performance test on the QPZZ-II gearbox dataset, as shown in Table 9. Consistent with the conclusion on the XJTU gearbox dataset, the proposed LWconvformer also exhibits excellent real-time performance on the QPZZ-II dataset, with a single-sample inference latency of only 3.681 ms and a processing speed of 271.6 FPS, which fully meets the latency requirements of real-time fault diagnosis in industrial scenarios. Compared with other models, MobileNet and ResNet18 achieve slightly lower inference latency but have significantly higher parameters and computational complexity. Models such as Conformer-NSE, MCSwin-T, and MobileNetV2 generally have an inference latency exceeding 5 ms, resulting in poor real-time performance. Overall, LWconvformer achieves a balance between diagnostic

accuracy and inference efficiency on different datasets, demonstrating good generalization and industrial deployment potential.

Table 9. Test results of real-time inference performance of comparison models in case 2

Model	Inference latency (ms)	FPS
LWConvFormer	3.681	271.6
CLFormer	4.534	220.5
Conformer-NSE	5.737	176.1
MCSwin-T	5.814	172.1
MobileNet	2.702	370.1
MobileNetV2	5.326	187.8
ResNet18	3.684	360.2
MK-ResCNN	4.104	235.2

4.2.3. Ablation study analysis

To further verify the generalization of the proposed module design, three groups of ablation experiments are conducted on the QPZZ-II bearing dataset, as listed in Table 10.

The model with DSMC and standard MSA achieves 93.37 % accuracy with significantly higher parameters and computational cost, which demonstrates the redundant computation and accuracy loss brought by traditional self-attention. The model with vanilla MC convolution and BSA reaches 93.47 % accuracy, along with sharp growth of parameter quantity and FLOPs, which verifies the lightweight multi-scale feature extraction advantage of the DSMC module.

The full LWConvFormer model obtains the optimal diagnostic accuracy of 94.01 %, as well as the minimum parameter of 0.322 M among all groups. The results reconfirm that the synergistic structure of DSMC and BSA achieves the optimal trade-off between diagnosis accuracy and model lightweight, which proves the favorable cross-dataset generalization of the proposed framework.

Table 10. The diagnostic accuracy and model complexity of each model in Case 2

Model structure	Ablation mode	Average accuracy (%)	Parameters (M)	FLOPs (M)
DSMC + standard MSA	Ablation_a	93.37	1.620	18.265
Standard MC + BSA	Ablation_b	93.47	1.237	38.781
Full LWConvFormer	Full	94.01	0.322	14.397

5. Conclusions

To tackle the problem of excessive model complexity arising from the multi-head self-attention mechanism in Transformers and cross-channel convolution in convolutional neural networks (CNNs), this paper introduces a lightweight fault diagnosis framework, LWConvFormer. This framework facilitates sample-adaptive feature extraction via a dynamic separable multi-scale convolution (DSMC) module and reduces computational complexity from quadratic to linear by integrating a broadcast self-attention (BSA) module.

Experiments conducted on the planetary gearbox and the QPZZ-II rotating mechanical test bench demonstrate the following: (1) The proposed method exhibits significant advantages over existing diagnostic techniques based on CNN and Transformer, particularly in terms of lightweight design and noise robustness; (2) The proposed framework reduces the number of parameters and computational load by a factor of 6 to 10 compared to mainstream methods, while increasing the training speed by nearly 7 times. The lightweight diagnostic framework presented in this paper offers an effective solution for deployment in industrial settings.

It should be noted that the datasets used in this study are collected under controlled laboratory conditions, and the performance under complex real industrial conditions – including sensor drift, temperature fluctuations, and non-stationary speed variations – requires further verification. Under such conditions, distribution shifts in vibration signals may lead to a certain degree of

performance degradation, especially when interference intensity exceeds the model's robustness boundary. To address this issue, future work will explore adaptive normalization and domain adaptation strategies to improve the model's generalization to real-world industrial data. Additionally, the current study focuses primarily on supervised learning scenarios with sufficient labeled data, while the scarcity of labeled fault samples remains a key challenge in practical applications. Future research will therefore focus on exploring the framework's applications in few-shot learning, interpretability analysis, and optimization under real industrial conditions to further enhance its practical value.

Acknowledgements

National Natural Science Foundation of China (Grant No. 52565011).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Peiyi Yang: conceptualization, supervision, writing-review and editing. Yizhe Song: methodology, software, formal analysis, writing-original draft. Tiantong Zhang: supervision, formal analysis, writing-review and editing.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] L. Zhang, K. Zhang, J. Liu, C. Wang, Y. Song, and S. Pan, "A bearing fault diagnosis method using bitmap image encoding and multimodal feature fusion," *Measurement*, Vol. 258, p. 119430, Oct. 2025, https://doi.org/10.1016/j.measurement.2025.119430

[2] L. Zhang et al., "A bearing fault diagnosis method with implementation of multiscale time-frequency feature fusion and bidirectional information interaction," *IEEE Sensors Journal*, Vol. 25, No. 13, pp. 23740–23756, Jul. 2025, https://doi.org/10.1109/jsen.2025.3568401

[3] Y. Wu, B. Tang, L. Deng, and Q. Li, "Distillation-enhanced fast neural architecture search method for edge-side fault diagnosis of wind turbine gearboxes," *Expert Systems with Applications*, Vol. 208, p. 118049, Jul. 2022, https://doi.org/10.1016/j.eswa.2022.118049

[4] J. Ou et al., "Accelerating long-context inference of large language models via dynamic attention load balancing," *Knowledge-Based Systems*, Vol. 333, p. 115018, Dec. 2025, https://doi.org/10.1016/j.knosys.2025.115018

[5] A. Vaswani et al., "Attention Is All You Need," *arXiv:1706.03762*, Aug. 2023, https://doi.org/10.48550/arxiv.1706.03762

[6] S. Tang, Y. Zhu, and S. Yuan, "Intelligent fault identification of hydraulic pump using deep adaptive normalized CNN and synchrosqueezed wavelet transform," *Reliability Engineering and System Safety*, Vol. 224, p. 108560, Aug. 2022, https://doi.org/10.1016/j.ress.2022.108560

[7] Y. Ding, M. Jia, Q. Miao, and Y. Cao, "A novel time-frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings," *Mechanical Systems and Signal Processing*, Vol. 168, p. 108616, Apr. 2021, https://doi.org/10.1016/j.ymssp.2021.108616

[8] B. Aslan, S. Balci, and A. Kayabasi, "Fault diagnosis in thermal images of transformer and asynchronous motor through semantic segmentation and different CNN models," *Applied Thermal Engineering*, Vol. 265, p. 125599, Jan. 2025, https://doi.org/10.1016/j.applthermaleng.2025.125599

[9] H. Fang et al., "CLFormer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions,"

IEEE Transactions on Instrumentation and Measurement, Vol. 71, pp. 1–8, 2021, https://doi.org/10.1109/tim.2021.3132327

[10] Z. Chen, J. Chen, S. Liu, Y. Feng, S. He, and E. Xu, "Multi-channel calibrated Transformer with shifted windows for few-shot fault diagnosis under sharp speed variation," *ISA Transactions*, Vol. 131, pp. 501–515, Apr. 2022, https://doi.org/10.1016/j.isatra.2022.04.043

[11] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mechanical Systems and Signal Processing*, Vol. 168, p. 108653, Dec. 2021, https://doi.org/10.1016/j.ymssp.2021.108653

[12] J. Deng, W. Jiang, Y. Zhang, G. Wang, S. Li, and H. Fang, "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*, Vol. 70, pp. 1–9, 2021, https://doi.org/10.1109/tim.2021.3091498

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv:1711.05101*, Jan. 2019, https://doi.org/10.48550/arxiv.1711.05101

[14] S. Han, H. Shao, J. Cheng, X. Yang, and B. Cai, "Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information," *IEEE/ASME Transactions on Mechatronics*, Vol. 28, No. 1, pp. 340–349, Aug. 2022, https://doi.org/10.1109/tmech.2022.3199985

[15] E. Elfatimi, R. Eryigit, and L. Elfatimi, "Beans leaf diseases classification using MobileNet models," *IEEE Access*, Vol. 10, pp. 9471–9482, Jan. 2022, https://doi.org/10.1109/access.2022.3142817

[16] L. Yong, L. Ma, D. Sun, and L. Du, "Application of MobileNetV2 to waste classification," *Plos One*, Vol. 18, No. 3, p. e0282336, Mar. 2023, https://doi.org/10.1371/journal.pone.0282336

[17] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, Vol. 12, No. 18, p. 8972, Sep. 2022, https://doi.org/10.3390/app12188972

[18] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 6, pp. 3797–3806, Jun. 2019, https://doi.org/10.1109/tii.2019.2941868



Peiyi Yang received his postgraduate degree, and now he works as a lecturer at Zhengzhou Railway Vocational and Technical College. His current research interests include urban rail vehicle technology, fault diagnosis, image acquisition and vocational education.



Yizhe Song is currently a postgraduate candidate pursuing his master's degree. His current research interest focuses on intelligent fault diagnosis.



Tiantong Zhang received his postgraduate degree, and now he works as an associate professor at Zhengzhou Railway Vocational and Technical College. His current research interests include urban rail vehicle technology, fault diagnosis and vocational education.