ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

A Dual-Channel Feature Fusion Network of MACNN and Informer for Gearbox Fault Diagnosis under Variable Operating Conditions

Zeyu Yang [1, a], Lijie Zhang [1, b *], Ze Xu [2, c], Linfeng Han [1, d], Haotian Guo [1, e], Jiacheng Xu [1, f]

[1] College of Electric Power, Inner Mongolia University of Technology, Hohhot, Inner Mongolia 010051;

[2] Inner Mongolia Today Internet Technology Co., Hinggan league, Inner Mongolia 137713.

[a] 20231800583@imut.edu.cn, [b*] zhanglijie@imut.edu.cn, [c] 2281478679@qq.com,

[d] 20231800587@imut.edu.cn, [e] 20231100460@imut.edu.cn, [f] 20231100445@imut.edu.cn

Abstract. To address the difficulty of extracting and identifying gearbox fault features under varying operating conditions in industrial environments, a gearbox fault diagnosis method based on a dual-channel feature extraction and fusion network combining a multi-scale convolutional neural network (MACNN) and an Informer encoder is proposed. First, variational mode decomposition (VMD) and the fast Fourier transform (FFT) are applied to preprocess the raw vibration signals to explore multi-scale characteristics of fault information. A local pyramid attention mechanism is then introduced, and the MACNN architecture is designed to extract local features, while the Informer encoder is employed to capture global features from gearbox vibration signals. Subsequently, a frequency-domain self-attention feature fusion module based on adaptive spectral filtering (FMFF) is constructed to integrate local and global features, and the fault classification results are obtained through attention-weighted pooling and a linear layer. The proposed method is validated on the MCC5-THU gearbox fault dataset, and experimental results demonstrate that, under identical preprocessing conditions, the proposed MACNN-Informer dual-channel feature fusion network consistently outperforms MACNN, CNN-LSTM, and 1D-CNN-Informer models. The average fault classification accuracy reaches 99.16%. These results verify the effectiveness and feasibility of the proposed method under variable operating conditions, indicating strong practical applicability.

Keywords: Gearbox; Compound Fault; Variable Operating Conditions; Attention Mechanism; Multi-Scale Convolutional Neural Network (MACNN); Informer; Feature Fusion.

1. Introduction

Gearboxes, as rotating component systems in mechanical equipment, are widely used in fields such as wind power generation, aerospace, and robotics [1,2]. In practical applications, gearboxes often operate under variable speed and variable load conditions, where complex excitations may induce faults in components such as gears and bearings. These faults can further evolve into compound faults with multiple fault couplings [3]. Therefore, the timely detection of gearbox faults under variable operating conditions is of great significance for ensuring the safe and reliable operation of mechanical equipment.

In recent years, to further improve the accuracy of gearbox fault diagnosis, researchers have attempted to fuse different deep learning models to jointly capture local and global features from vibration data. For example, Shixi Hou et al. [4] proposed a bearing fault diagnosis method based on the joint feature extraction of a Transformer and a residual neural network (ResNet). Experimental results demonstrated that the proposed fusion model achieved higher diagnostic accuracy in high-noise environments compared with single deep learning models such as Transformer, ResNet, and GRU. Zhang Long et al. [5] constructed a parallel CNN–GRU dual-channel network, in which the CNN was used to extract important local spatial features from gearbox vibration signals, while the GRU efficiently captured global temporal features. The extracted features were concatenated and fed into a Softmax classifier for fault classification,

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21
achieving a classification accuracy of 99.7% on the Zhejiang University gearbox fault dataset. In 2025, Ma Shuangbao et al. [6] proposed a fault diagnosis method based on a 1D-CNN–Informer architecture. In this approach, vibration signals were preprocessed using the Fast Fourier Transform (FFT) and variational mode decomposition (VMD), and a dual-branch feature extraction framework was constructed using 1D-CNN and the Informer network. The Informer[7] effectively avoids gradient vanishing or explosion problems and overcomes the limitations of long-distance temporal modeling, demonstrating superior capability in capturing long-term dependencies. By employing a multi-head attention mechanism for feature fusion and a Softmax classifier for fault identification, the method achieved higher diagnostic accuracy than CNN–LSTM, TimesNet, and single Informer models.

Based on the above analysis, to address the challenges of difficult fault feature extraction and low recognition accuracy of compound faults under time-varying operating conditions, this paper proposes a gearbox fault diagnosis method based on a MACNN–Informer dual-channel feature fusion network. First, VMD and FFT are applied to preprocess gearbox vibration data to construct time–frequency feature matrices. Subsequently, a MACNN based on a local pyramid attention mechanism is designed to extract local fault features through the joint modeling of multi-scale convolution and local attention. Meanwhile, the Informer encoder is employed to extract global temporal features from vibration signals. Finally, a multi-head frequency-domain self-attention feature fusion module is proposed to integrate local and global features, thereby enhancing the feature extraction capability and fault recognition performance of the model. To verify the effectiveness of the proposed method under time-varying operating conditions, experiments are conducted on the MCC5-THU gearbox multimode fault dataset.

2. Design of the MACNN Network

2.1 Structure of the MCNN Network

MCNN [8] is a neural network with a multi-branch parallel architecture. In each branch, convolutional layers with different kernel sizes are employed. By performing convolution and pooling operations at multiple scales, the network is able to perceive local features of the input sequence at different temporal scales. The multi-scale features extracted from all branches are then concatenated along the channel dimension to construct a unified feature representation.

The data processing procedure of each branch can be expressed as:

$$f_{\mathrm{k}}(x) = MaxPool\left(\sigma\left(x * W_{\mathrm{k}} + \mathrm{b}_{\mathrm{k}}\right)\right) \tag{1}$$

In this article, $f_k(x)$ denotes the convolution and pooling operations of the k-th branch, $\mathrm{Maxpool}(\cdot)$ represents the max-pooling operation, $\sigma(\cdot)$ is the sigmoid activation function, W_k denotes the convolution kernel of the k-th branch, $*$ denotes the convolution operation, and $b_k$represents the corresponding bias term.

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

2.2 The Basic Principle of the Local Pyramid Attention Mechanis

The Local Pyramid Attention (LPA) mechanism is an attention mechanism proposed by Hao Du et al. [9] in 2024 to further enhance the model's capability for local feature learning. The structure of LPA is illustrated in Fig. 1. It consists of a Spatial Attention (SA) module and a Channel Attention (CA) module. The SA module is designed to identify the time segments that are more critical for detection and recognition tasks within the sequence, while the CA module is used to evaluate the relative importance of features across different channels.For an input feature X, the attention weights of the SA module are computed as:

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

$$A_S = \sigma\left(\xi\left(Concat\left(P_{avg}\left(X\right), P_{max}\left(X\right)\right)\right)\right) \tag{2}$$

In this article, P_{avg} and P_{max} denote adaptive average pooling and adaptive max pooling along the channel dimension, respectively, $Concat$ denotes the concatenation operation, and $\xi(\cdot)$ represents a one-dimensional convolution operation with a kernel size of 9.

The attention weights of the CA module are calculated as:

$$A_C = \sigma\left(\mathrm{FC}_2\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\mathrm{Pool}_{\text{avg}}\left(X\right)\right)\right)\right)\right) \tag{3}$$

In this article, $\mathrm{Pool}_{\text{avg}}$ denotes global average pooling along the channel dimension, FC represents a fully connected layer, and ReLU denotes the activation function.

For an input sequence X, the LPA attention mechanism processes the sequence by dividing it into i hierarchical levels. Taking $i = 2$ as an example, the structure of LPA is shown in Fig. 1.

The first level P_1 operates directly on the entire sequence for global attention modeling, where global attention weight sequences are generated successively through the CA module and the SA module. The second level P_2 divides the input sequence along the temporal dimension into 2^2 equal-length subsequences for more fine-grained local attention modeling. Each subsequence is first processed by the CA module to learn a channel weight vector, which is then multiplied with the subsequence itself to obtain channel-weighted features, as expressed by the following equation:

$$W_{\text{Cj}} = \mathrm{f_j} * \mathrm{w}_C \tag{4}$$

In this article, f_j denotes the input subsequence, W_C represents the matrix corresponding to the CA module, and W_{Cj} denotes the channel-weighted feature.

Subsequently, the SA module is applied to W_{Cj} to perform weighted computation, generating the corresponding attention weight sequence W_{Sj}, as expressed by the following equation:

$$W_{Sj} = W_{\text{Cj}} * \mathrm{w}_S \tag{5}$$

In this article, W_S denotes the matrix corresponding to the SA module.

Subsequently, the four local attention features are concatenated to generate the enhanced local features, which constitute the attention-enhanced feature sequence output by the second level P_2. The resulting feature sequence has the same dimensionality as the original input. Finally, the enhanced global features output by P_1 and the enhanced local features output by P_2 are fused through element-wise addition to form the final attention weight sequence.

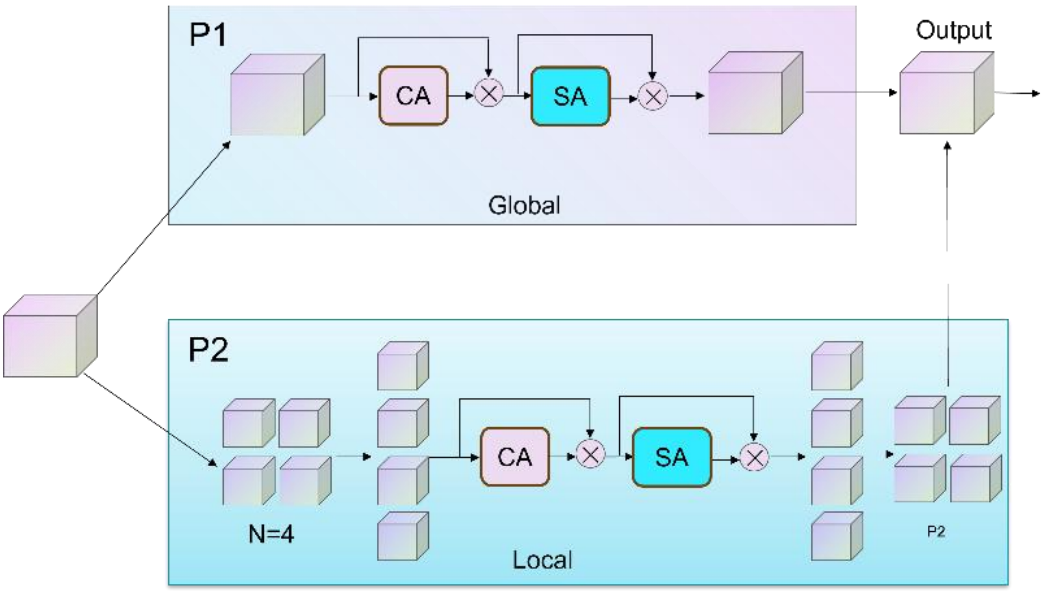


Fig. 1. Local Pyramid Attention Mechanism Structure

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

2.3 The Construction of the MACNN Model

To address the difficulty of effectively extracting fault features from gearbox vibration signals under time-varying operating conditions, the MCNN is further improved and an attention-based multi-scale convolutional neural network (MACNN) is designed. The structure of the proposed MACNN is shown in Fig. 2. It mainly consists of a wide convolution layer, multi-scale convolutional branch modules, and an attention mechanism. The main network parameters of the model are listed in Table 1.

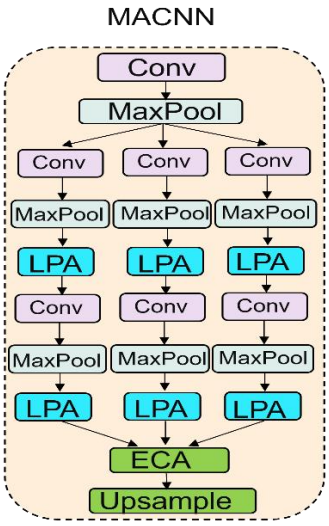


Figure 2. Structure of MACNN

Table 1. Three Scheme comparing

Layer name	Kernel size	Output channels	Stride
Wide convolution layer	32 × 1	32	2
Pooling layer	2 × 1	32	2
Multi-scale convolution layer 1	1 × 1 / 5 × 1 / 9 × 1	64 / 64 / 64	1 / 1 / 1
Multi-scale pooling layer 1	2 × 1 / 2 × 1 / 2 × 1	64 / 64 / 64	2 / 2 / 2
Attention layer 1	–	64 / 64 / 64	–
Multi-scale convolution layer 2	5 × 1 / 7 × 1 / 9 × 1	128 / 128 / 128	1 / 1 / 1
Multi-scale pooling layer 2	2 × 1	128 / 128 / 128	2 / 2 / 2
Attention layer 2	–	128 / 128 / 128	–
Upsampling	–	384	–
Channel compression	1 × 1	128	1

3. Literature References

The Informer encoder [7] mainly consists of a multi-head probabilistic sparse self-attention mechanism module and a distilling module, and its overall structure is illustrated in Fig. 3.The probabilistic sparse self-attention mechanism achieves effective feature representation by suppressing redundant interactions and highlighting key dependency relationships, thereby significantly reducing the computational complexity of long-sequence modeling. This attention mechanism retains only the top k Keys that have the greatest influence on the current Query, enabling the model to focus on critical segments of the input time series. The corresponding computation formula is given as follows:

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\bar{\boldsymbol{Q}}\boldsymbol{K}^{\text{ú}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{6}$$

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

In this article, Q denotes the query vector, K denotes the key vector, and V denotes the value vector; K_T represents the Keys obtained after random sampling; $\bar{Q}$ denotes the probability sparse matrix obtained through sparsification; Softmax is the activation function; and d denotes the dimensionality of the input vector.

Since the output features obtained after the probabilistic sparse self-attention mechanism contain a large amount of redundant information, the distilling operation is introduced to compress the temporal dimension of the features, thereby reducing the computational burden of subsequent layers. The distilling module consists of a one-dimensional convolutional layer and a max-pooling layer. The mathematical formulation of the distilling operation from the i-th layer to the $(i+1)$-th layer is given as follows:

$$X^{(i+1)} = \text{MaxPoollD}\left(\sigma\left(\text{Conv1D}\left(X^{(i)}\right)\right)\right) \tag{7}$$

In this article, MaxPool1D denotes the max-pooling operation along the temporal dimension, and Conv1D denotes the one-dimensional convolution operation.

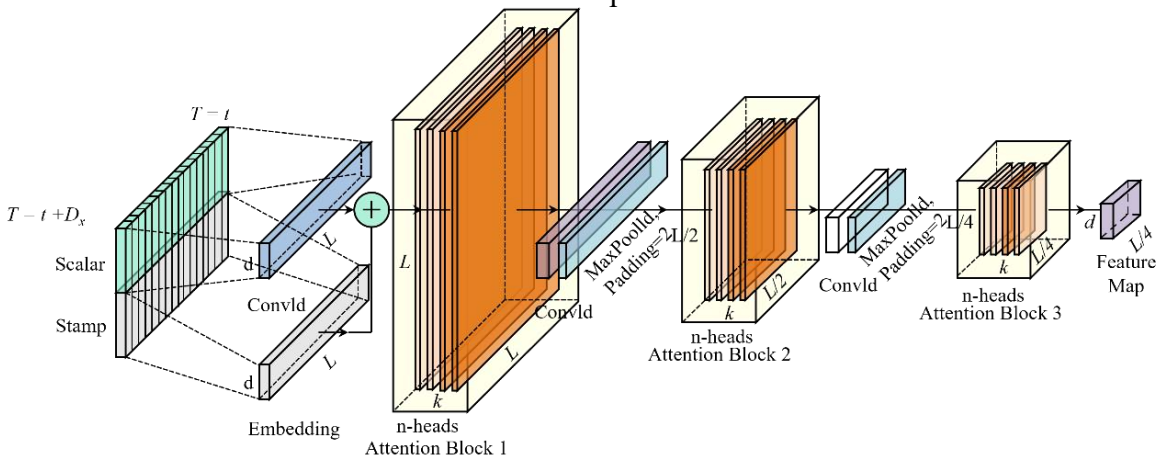


Fig. 3 Informer encoder structure

4. MACNN-Informer Dual-Channel Feature Fusion Network Design

4.1 Multi Head Spectral Attention

The Multi-Head Spectral Attention (MHSA) mechanism was proposed by Jacob et al. [12] in 2025. Its core idea is to transform sequences from the time domain to the frequency domain and highlight key frequency bands of the input sequence through adaptive spectral filters, thereby replacing the explicit computation of attention weight matrices. The features are then remapped back to the time domain via a nonlinear activation function and the inverse fast Fourier transform (iFFT). In this way, MHSA can model long-range dependencies at low computational cost while further enhancing feature representation capability. The structure of MHSA is illustrated in Fig. 4 and consists of the following four steps:

(1) The input sequence X is first normalized and then divided into *h*attention heads.

(2) Hybrid frequency-domain feature extraction. The FFT is applied to the input sequence to obtain global spectral information. Meanwhile, the input sequence is segmented into multiple overlapping Hamming windows, and the FFT of each windowed segment is computed to obtain local frequency information. The global spectral information and local spectral information are then concatenated along the feature dimension to form the hybrid frequency-domain feature F .If you follow the “checklist” your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

(3) A global context vector is computed and fed into a lightweight multi-layer perceptron (Multi-Layer Perceptron, MLP) to generate scaling parameters Δs and bias parameters Δb for each

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21
attention head. The parameters $\Delta s$and $\Delta b$are then combined with the scaling and bias parameters of a learnable base filter through element-wise multiplication and addition, respectively, to construct an adaptive spectral filter. The hybrid frequency-domain features are linearly modulated by the adaptive spectral filter to obtain the modulated frequency-domain features $\mathbf{F}^{'}$. The above process can be expressed as follows:

$$\mathrm{c} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{X}_i \tag{8}$$

$$W = W_{\text{base}} \odot \left(1+\Delta s\right), b = b_{\text{base}} + \Delta b \tag{9}$$

$$F' = F \odot W + b \tag{10}$$

In this article, n denotes the length of the input sequence, c represents the global context vector, W and b are learnable parameters of the filter, and $\odot$denotes element-wise multiplication.

(4) Time-domain fusion feature extraction. A complex-valued activation function based on magnitude modulation is applied to enhance the modulated frequency-domain features $\mathbf{F}^{'}$, and the enhanced frequency-domain features are then transformed back to the time domain via the inverse fast Fourier transform (IFFT). The time-domain features output by each attention head are concatenated to obtain the fused time-domain feature representation. The above process can be expressed as follows:

$$F_{\text{act}} = \phi\left(F'\right) = F' \cdot \frac{\mathrm{GELU}\left(\left|F'\right|\right)}{\left|F'\right| + \dot{o}} \tag{11}$$

$$x_{\text{filtered}}^{(m)} = \mathrm{F}^{-1}\left(F_{\text{act}}^{(m)}\right), \quad m = 1,\ldots,h \tag{12}$$

$$Y = \mathrm{concat}\left(\left[x_{\text{filtered},1}, x_{\text{filtered},2}, \ldots x_{\text{filtered,h}}\right]\right) \tag{13}$$

In this article, F_{act} denotes the enhanced frequency-domain feature, $\phi(\cdot)$ represents the complex-valued activation function based on magnitude modulation, GELU denotes a smooth nonlinear transformation applied to the magnitude component, $\epsilon$is a numerical stability parameter used to prevent division by zero, $\mathrm{x}_{\text{filtered}}$denotes the time-domain feature output of a single attention head, $\mathcal{F}^{-1}$denotes the IFFT operation, and Y represents the fused time-domain feature.

4.2 Multi-Head Frequency Self-Attention Feature Fusion Module Design

Since vibration signals usually contain rich and structured frequency-domain information, in which fault-related features are often manifested prominently in the frequency domain, the time-domain fusion feature extraction strategy of the multi-head spectral attention mechanism is adopted to design a Multi-Head Frequency Self-Attention Feature Fusion module (Frequency Multi-Head Attention Feature Fusion, FMFF). The proposed FMFF is used to fuse features extracted by different networks while emphasizing their frequency-domain characteristics. The overall structure of FMFF is illustrated in Fig. 4, and the specific procedure is described as follows.

First, the features extracted by different networks are aggregated to obtain a hybrid feature $\mathrm{x} \in \mathbb{R}^{B\times L\times D}$,where Bdenotes the number of samples, $L$is the sequence length, and $D$is the embedding feature dimension. The hybrid feature $\mathbf{x}$is then split into $h$attention heads. For each attention head, the input sequence is transformed into the frequency domain by applying the FFT, yielding the corresponding frequency-domain features, The above process can be expressed as follows:

$$x = x_{\text{Local}} + x_{\text{Global}} \tag{14}$$

$$x \to x_h \in \mathrm{R}^{B\times h\times L\times D_h}, F = \mathrm{F}\left(x_h\right) \tag{15}$$

In this article, $D_h$denotes the feature dimension of each attention head, $\mathcal{F}(\mathbf{x})$represents the FFT operation, and $\mathbf{x}_{\text{Local}}$ and $\mathbf{x}_{\text{Global}}$ denote the local features and global features extracted by the preceding networks, respectively.

Next, the multi-head attention input sequences are processed according to Steps (3)–(4) of the multi-head spectral self-attention mechanism described in Section 4.1. To enhance model stability

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21
and preserve the original temporal information, a residual connection is introduced in the frequency-domain enhancement module. As a result, the module output contains not only the original hybrid features but also the correction information obtained from frequency-domain enhancement. This design enables more stable utilization of frequency-enhanced features and effectively alleviates overfitting.

Finally, an attention-weighted pooling mechanism [13] is applied to the fused time-domain features **Y**to dynamically adjust the importance of data at different time steps. The attention-pooled features are then fed into a classifier to complete fault pattern recognition.

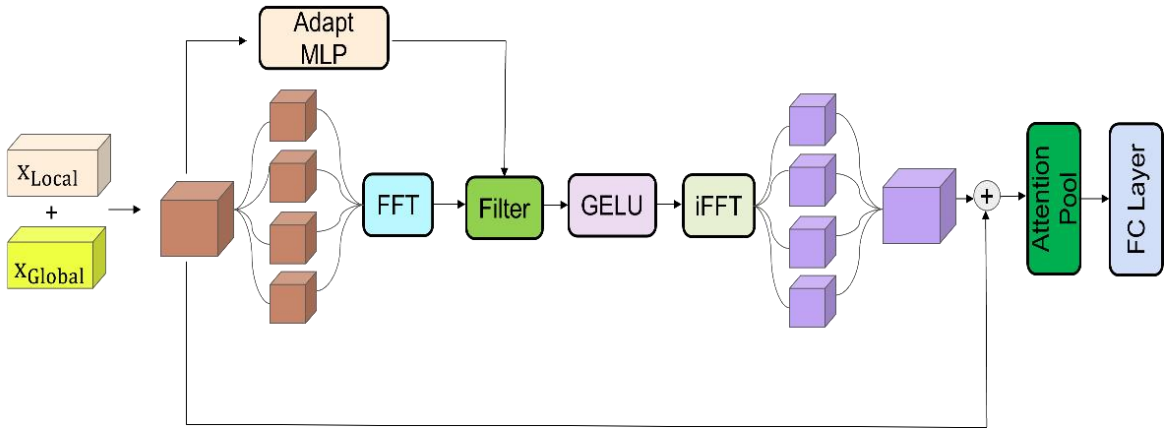


Fig. 4. Multi-Head Frequency Self-Attention Feature Fusion Module Structure

4.3 Architecture of the MACNN–Informer Dual-Channel Feature Fusion Network

To simultaneously explore multi-scale time–frequency features and global temporal dependency information in gearbox vibration signals and to improve fault discrimination accuracy, a MACNN–Informer dual-channel feature fusion network is proposed. The overall architecture of the network is shown in Fig.5 and consists of three main components: data preprocessing, feature extraction networks, and a feature fusion module.

(1) Data preprocessing: To enhance the time–frequency representation capability of the input data, variational mode decomposition (VMD) is first applied to the input signals, and the fast Fourier transform (FFT) is simultaneously performed to obtain frequency-domain information. The main intrinsic mode functions and the spectral magnitude vectors are then concatenated along the channel dimension to construct multi-channel one-dimensional data, which serve as the input to the feature extraction networks.

(2) Feature extraction networks: To capture both local texture details and global evolution trends of the input data, the MACNN designed in Section 2.3 is employed to extract local features, while the local pyramid attention mechanism is incorporated to further enhance the modeling capability of key details. In parallel, Global features are extracted using the Informer encoder.

(3) Feature fusion module: To achieve interaction and joint modeling of different feature representations, the FMFF module designed in Section 4.2 is utilized to fuse the global and local features extracted by the feature extraction networks, thereby producing the final fault discrimination results.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

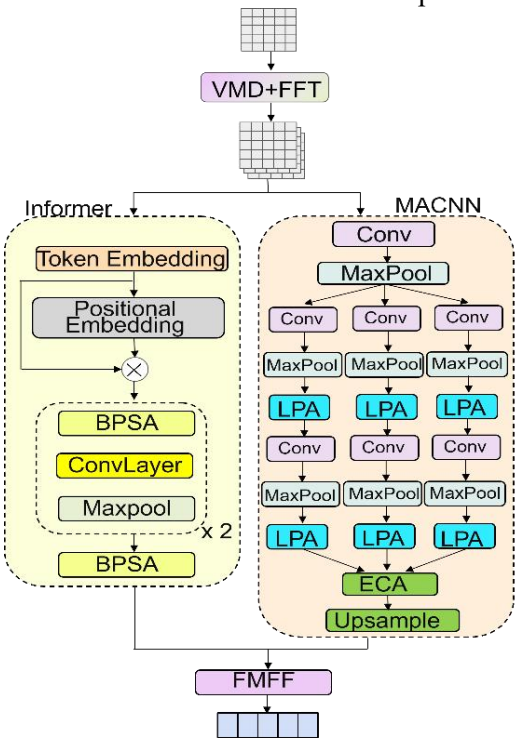


Fig. 5. MACNN- Informer Dual-Channel Feature Fusion Network Structure

5. Experiments

5.1 Dataset and Experimental Data Selection

Experiments are conducted using the MCC5-THU gearbox multimode fault dataset [14], which was publicly released in 2024 by China Metallurgical Group Corporation (MCC Group) and Tsinghua University. The experimental test rig used to construct the dataset is shown in Fig. 6.

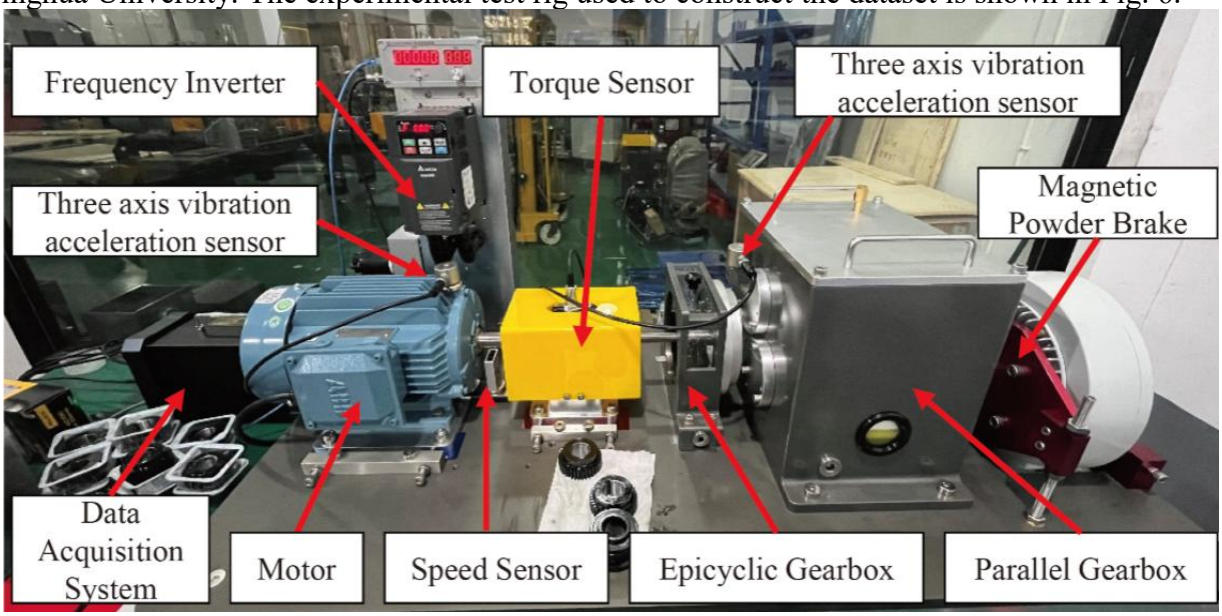


Fig. 6. Gearbox Fault Diagnosis Experimental Platform

In practical gearbox operation, load fluctuations are more common and almost ubiquitous. Therefore, the experimental data are collected under time-varying load conditions, where the motor input shaft rotates at a constant speed of 1000 rpm, while the output shaft torque of the gearbox varies over time within the range of 0–10 Nm.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

The vibration acceleration data along the x-direction of the intermediate shaft are selected. A total of 768,000 data points are recorded over 60 s. The time window length is set to 2048 points per sample, and a 50% overlap is maintained between consecutive samples, corresponding to a stride of 1024 data points. As a result, 748 samples are generated for each fault condition, and 5984 samples are obtained in total for eight fault conditions. All samples are divided into a training set and a test set at a ratio of 4:1.

5.2 Model Training

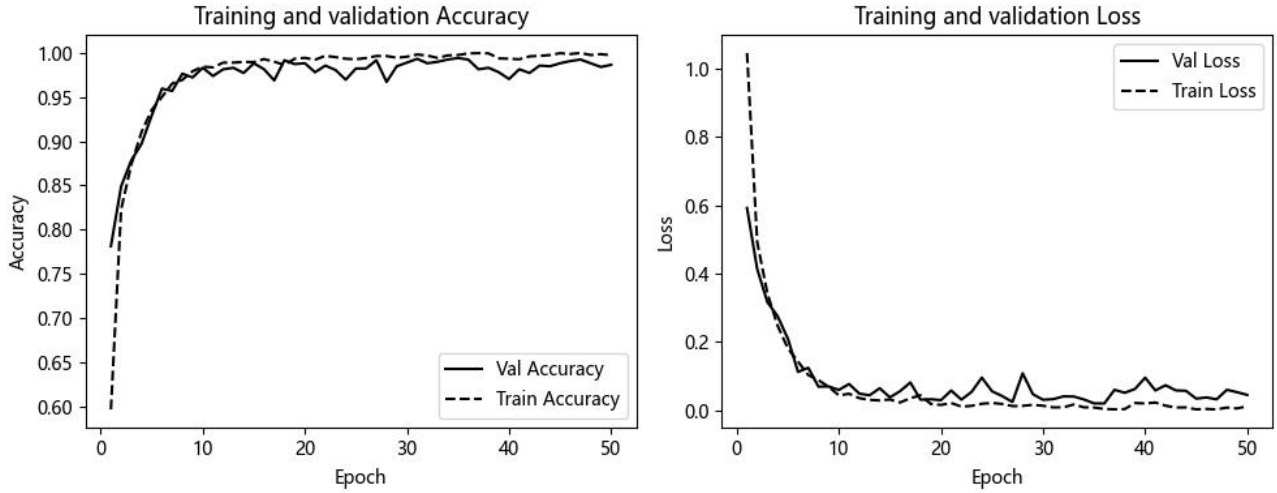
The experimental platform is a self-configured computer with a 64-bit Windows 11 operating system, equipped with a 13th Gen Intel® Core™ i7-13650HX processor, an NVIDIA® GeForce RTX™ 4060 Laptop GPU, and 16 GB of RAM. The program is developed based on the PyTorch framework using Python 3.9. During training, the Adam optimizer is employed with an initial learning rate of 0.0003, and the cross-entropy loss function is adopted. The model is trained for 50 epochs using a fixed learning rate strategy.

The MACNN–Informer dual-channel feature fusion network is trained on the training set, and the fault recognition accuracy and loss curves during the training process are shown in Fig. 7. As can be observed from Fig. 7, both the fault recognition accuracy and the loss function tend to converge after the 10th epoch, and the final training accuracy reaches 99.9%.

After 50 training epochs, the trained model is further evaluated on the test set. Fig. 8 presents the confusion matrix obtained from the first test run. The average classification accuracy of the eight fault types exceeds 99%, indicating that the MACNN–Informer network exhibits strong discriminative capability for fault features. The fault types corresponding to C1–C8 in the figure are listed in Table 2.

Table 2. Types of Faults

Label	Fault type	Training samples	Test samples
C1	Gear crack	596	148
C2	Broken tooth and bearing inner race fault	596	148
C3	Broken tooth and bearing outer race fault	596	148
C4	Gear pitting	596	148
C5	Broken tooth	596	148
C6	Healthy	596	148
C7	Gear wear	596	148
C8	Missing tooth	596	148



(a)Training Accuracy (b)Training Cross-Entropy Loss

Fig. 7. Curves of Accuracy and Loss during the Training Process

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

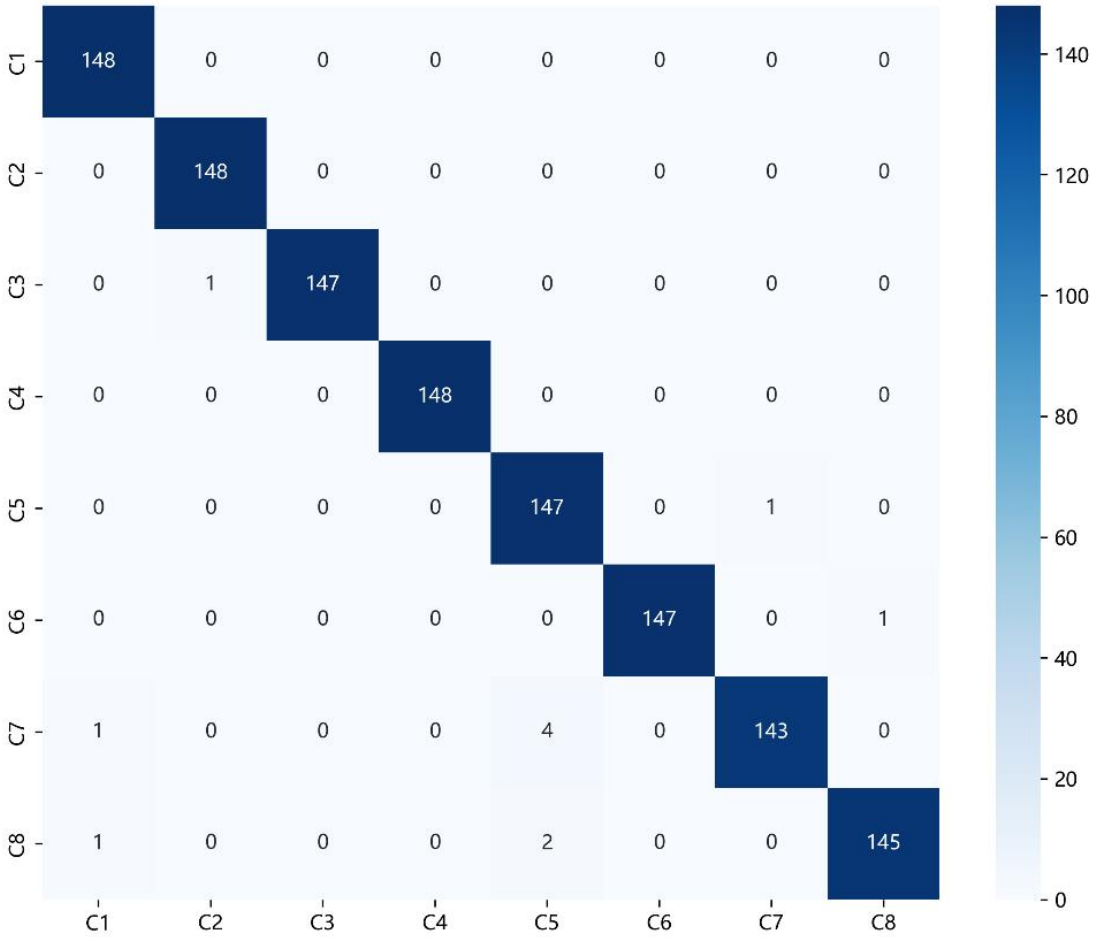


Fig. 8. Classification Results of the MACNN Informer Model

5.3 Comparative Experiments of Fault Diagnosis Algorithms

To verify the superiority of the proposed MACNN–Informer dual-channel feature fusion network, comparative experiments are conducted against several mainstream fault diagnosis methods, including VGG11, MACNN [15], a serial CNN–LSTM model [16], a parallel CNN–GRU model [5], and 1D-CNN–Informer [6]. For a fair comparison, all input data are preprocessed using VMD and FFT, and all models are trained with the same training parameters for 50 epochs. The diagnostic accuracy curves of different models during training are shown in Fig. 9. Subsequently, 10 fault diagnosis trials are performed on the test set. The effectiveness of the proposed gearbox fault diagnosis method is quantitatively evaluated using the average values of Precision, Recall, F1-score, and Accuracy [6] across these trials, with detailed results presented in Table 3.

As shown in Fig. 9, the MACNN–Informer model outperforms the other models in terms of both convergence speed and classification accuracy. Furthermore, as indicated in Table 3, the proposed MACNN–Informer model achieves an average diagnostic accuracy of 99.16% under time-varying operating conditions, representing an improvement of at least 1.57% compared with the other methods.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21

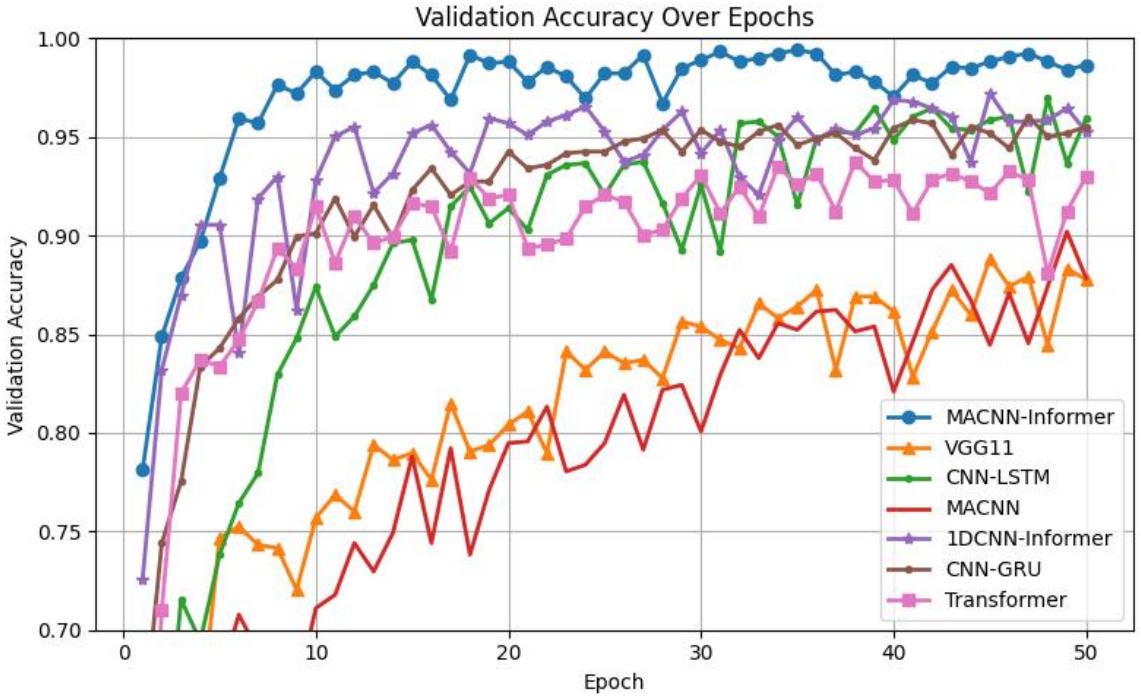


Fig. 9. Fault Diagnosis Accuracy Curves of Different Models during Training

Table 3. Results of Comparative Experiments

Model	Precision	Recall	F1-score	Accuracy/%
VGG11	89.15	87.78	87.97	87.84
MACNN	88.33	87.74	87.84	87.75
CNN-GRU	97.58	97.52	97.50	97.59
CNN-LSTM	96.11	95.95	95.96	96.03
1DCNN-Informer	96.81	96.73	96.55	96.74
MACNN-Informer	99.14	99.16	99.15	99.16

6. Conclusion

A MACNN–Informer dual-channel feature fusion network is proposed for gearbox fault diagnosis under time-varying operating conditions. The network combines an LPA-based MACNN for local feature extraction with an Informer encoder for global temporal modeling, and introduces a multi-head frequency self-attention fusion module to enhance fault-related frequency characteristics. Experimental results on the MCC5-THU dataset demonstrate that the proposed method achieves an average diagnostic accuracy of 99.16%, outperforming several state-of-the-art models and verifying its effectiveness for compound fault diagnosis under variable operating conditions.

Acknowledgements

This work was supported by the Inner Mongolia Science and Technology Program (No. 2023YFSW0015).

References

[1] Cédric Peeters, Jérôme Antoni, Jan Helsen. Blind filters based on envelope spectrum sparsity indicators for bearing and gear vibration-based condition monitoring. Mechanical Systems and Signal Processing, 2020, 138.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P21
[2] XIONG Xin, ZHENG Shaoshuai, HE Jun, YANG Shixi. Gearbox fault identification based on improved One-Dimensional convolutional neural network and similarity measure function enhancement. Journal of Vibration, Measurement and Diagnosis, 2022, 42, (2): 242-248, 404

[3] XIONG Fei, LI Haiping, WANG Long, et al. Fault diagnosis strategy for mechanical equipment under variable working conditions based on MFCC-DBN. Journal of Mechanical & Electrical Engineering, 1-9[2025-09-10].

[4] Han, Shuzhen,Li, Jianfei,Pang, Ke , et al.Bearing fault diagnosis using joint features extraction with multi-scale residual convolutional neural network and transformer. Measurement Science and Technology, 2025, 36, (5): 056108-056108.

[5] ZHANG Long, ZHEN Canzhuang, YI Jianyu , et al. Dual-channel feature fusion CNN - GRU gearbox fault diagnosis.Journal of Vibration and Shock, 2021, 40(19): 239-245+294.

[6] Shuangbao Ma,Songjie Shi,Yapeng Zhang , et al.A High-precision method for detecting rolling bearing faultis in unmanned aerial vehicle based on improved 1DCNN-Informer model. Measurement, 2025, 256.

[7] Zhou, Haoyi,Zhang, Shanghang,Peng, Jieqi , et al.Informer: beyond efficient transformer for long sequence time-series forecasting. //35th AAAI Conference on Artificial Intelligence / 33rd Conference on Innovative Applications of Artificial Intelligence / 11th Symposium on Educational Advances in Artificial Intelligence. 2021: 11106-11115.

[8] Zhang, Yingying,Zhou, Desen,Chen, Siqin , et al. Single-image crowd counting via multi-column convolutional neural network. //2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 589-597.

[9] Du, Hao,Wang, Jiazheng,Liu, Min , et al.SwinPA-Net: swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, (4): 5355-5366.

[10] Zhang, Wei,Peng, Gaoliang,Li, Chuanhao , 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): 42.

[11] Wang, Qilong,Wu, Banggu,Zhu, Pengfei , et al. ECA-Net: efficient channel attention for deep convolutional neural networks. //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 11531-11539.

[12] Fein-Ashley, Jacob,Gupta, Neelesh,Kannan, Rajgopal , et al. SPECTRE: an FFT-based efficient drop-in replacement to self-attention for long contexts. arXiv, 2025.

[13] Takeshi D. Itoh,Takatomi Kubo,Kazushi Ikeda. Multi-level attention pooling for graph neural networks: Unifying graph representations with multiple localities. Neural Networks, 2022, 145, 356-373.

[14] Shijin Chen,Zeyi Liu,Xiao He , et al. Multi-mode fault diagnosis datasets of gearbox under variable working conditions. Data in Brief, 2024, 54.

[15] ZHANG Xijun, SHANG Ji-yang, YU Guangjie. et al.Bearing fault diagnosis based on attention for multi-scale convolutional neural network. Journal of Jilin University (Engineering and Technology Edition), 2024, 54(10): 3009-3017.

[16] CHEN Yongzhan, QU Jianling, WANG Xiaofei, et al. Rolling bearing fault diagnosis based on CNN-LSTM with time sequential memory enhancement. Noise and Vibration Control, 2025, 45(01): 105-111.