https://doi.org/10.37965/jdmd.2026.1840

Time-Frequency Dual-Stream Fusion Network for Bearing Fault Diagnosis under Variable Speed Condition

Jinrui Wang[1], Zhu Chen[1], Huadong Du[2], Yufeng Xu[2], Zongzhen Zhang[1*], Baokun Han[1]

[1]College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, China;
[2]AUCMA Company Limited, Qingdao 266000, China;
*Corresponding author: zhzz18@126.com

Received 7 August , 2026; Revised 8 September,2026; Accepted 2 November, 2026; Published online 3 November , 2026.

Abstract: Fault diagnosis in mechanical systems under variable speed condition has long been limited by single-modality signal analysis methods, which often result in incomplete feature characterization and fail to fully reflect the true health status of equipment. Time-domain and frequency-domain signals contain a wealth of redundant and complementary fault information. However, how to efficiently extract and fuse these two types of heterogeneous signals remains a research challenge in the field of rotating machinery fault diagnosis. To address this issue, this paper proposes a rolling bearing fault diagnosis method based on a time-frequency dual-stream fusion network, which relies on time-frequency dual-stream feature fusion technology to achieve comprehensive and robust extraction of bearing fault features. The proposed method constructs a parallel network architecture with time-domain and time-frequency-domain branches. It introduces a bidirectional cross-attention mechanism between the time-domain and time-frequency-domain to achieve bidirectional guidance and interactive enhancement of the two types of features. Furthermore, through a cross-domain attention fusion strategy, it accomplishes the deep integration and complementarity of time-frequency heterogeneous features, thereby strengthening the cross-domain representation capability of fault-sensitive features. This approach not only preserves the transient impact information of time-domain signals but also takes into account the frequency distribution patterns of time-frequency-domain signals. The network also incorporates regularization strategies such as label smoothing and cosine-annealed learning rates, effectively enhancing the model's generalization ability and diagnostic stability. Finally, the effectiveness of the proposed method was validated through rolling bearing fault diagnosis experiments under variable speed condition.

Keywords: Rolling bearing; fault diagnosis; time-frequency dual streams; bidirectional cross-attention; feature fusion; variable speed condition

1. Introduction

In mechanical machinery, bearings serve as essential elements for sustaining rotary motion and carrying operational forces [1,2]. Nevertheless, these components are quite prone to defects arising from extended usage or severe environmental conditions [3], possibly resulting in costly manufacturing interruptions and substantial financial impacts [4,5]. Hence, implementing reliable and precise diagnostic techniques for bearings is crucial [6]. Motivated by progress in AI and data-centric strategies, smart fault identification has emerged as a prominent area within condition surveillance of bearings [7]. These advanced systems, which incorporate real-time data acquisition and detailed signal processing, enable the early recognition of deterioration indicators with improved accuracy [8].

Some existing research focuses on time-domain analysis, characterizing fault impacts and transient changes through raw vibration waveforms or statistical features [9]. Other studies adopt a frequency-domain perspective, utilizing spectral analysis to extract periodic structures and characteristic frequency information embedded in the signals [10]. Different signal characterization methods each have their own advantages under specific operating conditions, providing an important theoretical foundation for subsequent exploration of collaborative modeling and fusion analysis of time-domain and frequency-domain information[11]. Tian et al. [12] proposed a CNN-LSTM model based on a hybrid particle swarm optimization algorithm to improve diagnostic accuracy. Liu et al. [13] combined Wavelet Packet Decomposition (WPD) with CNN to propose an improved Recurrence Plot (RP) method, which reduces redundancy and enhances diagnostic accuracy

through multiscale asymmetric Multi-scale Asymmetric Recurrence Plot(MARP). Lu et al. [14] proposed a method based on Gramian Angular Field (GAF) coding and an improved ResNet50 model to achieve efficient and accurate diagnosis under big data conditions. Zhang et al. [15] converted vibration data into two-dimensional images using Continuous Wavelet Transform(CWT), utilized VAE-GAN to augment samples, and trained a CNN, demonstrating superior performance in fault classification.Mohiuddin et al. [16] optimized a pre-trained CNN model to reduce parameters and computation time, thereby improving diagnostic performance under various load and noise conditions. Kim et al. [17] proposed the TF-MDA model, which extracts time-frequency domain features via a parallel CNN-1D, combined with physics-driven preprocessing and attention modules, to enhance classification accuracy and robustness in high-noise environments.Zhao et al. [18] proposed an improved prototype-based fault diagnosis model designed to address the small-sample problem, further enhancing the robustness and discriminative power of fault diagnosis. Vaswani et al. [19] proposed a Transformer architecture based on self-attention mechanisms, which overcomes the limitations of CNNs and RNNs and has achieved significant success in the field of natural language processing. Variants such as Vision Transformer, and SwinTransformer [20,21] have further driven the widespread adoption of Transformers. Xu et al. [22] addressed the challenge of early fault detection in rolling bearings by proposing a diagnostic method based on a 1D-Transformer encoder, which directly inputs raw data to achieve end-to-end diagnosis and outperforms traditional methods in terms of accuracy, computational complexity, and noise robustness.

In industrial rotating machinery, rolling bearings mostly operate under variable speed condition, which generates non-stationary vibration signals with drifting fault characteristic frequencies and smeared spectral components. Single time or frequency domain analysis fails to capture stable fault signatures under speed fluctuations,.The strict cross-domain scenario of “training on discrete fixed-speed data and testing on entirely unseen variable-speed data” is a challenging open problem. In this work, training and test samples are segmented from continuously-changing-speed signals within overlapping speed ranges, and the generalization performance for rotational speeds completely outside the sampled interval is not verified in current experiments. However, existing deep learning-based fault diagnosis methods typically implicitly assume that the training and test data follow the same or similar distributions [23]. In practical engineering applications, however, due to factors such as variations in the frequency of fault types, limited data acquisition conditions, and diverse operating conditions, it is often difficult to ensure consistency in the scale and categorical composition of different fault samples [24]. While samples for some common faults are relatively abundant, the number of samples for certain fault types is limited due to their low occurrence probability or high acquisition cost. In cross-device or cross-operating-condition scenarios, there may also be discrepancies between the fault categories covered during the training phase and the fault types that actually occur during operation [25]. Such discrepancies in sample distribution directly affect the model’s feature learning effectiveness and diagnostic performance, and have gradually become a significant factor limiting the engineering application of intelligent fault diagnosis methods [26].

In the field of fault diagnosis, attention mechanisms have likewise gained broad adoption. For instance, Yang et al. [27] introduced a diagnosis approach that integrates a signal Transformer with attention modules to accomplish effective feature learning through partitioning and encoding raw univariate vibration data. Li et al. [28] presented the Multi-scale Convolution Attention Mobile Deep Network(MCAMDN) architecture, merging multiscale features, depthwise separable convolutions, and attention modules for tackling classification precision problems in noisy environments. Chen et al. [29] developed a balanced conditional CNN-Transformer generative adversarial network that combines subtle data augmentation with the Synthetic Minority Oversampling Technique(SMOTE) technique, thereby strengthening diagnostic accuracy and robustness under limited and class-skewed sample distributions. Furthermore, Fang et al. [30] put forward the lightweight SANet framework for fault detection, which blends enhanced convolutional blocks with a Transformer module to attain superior accuracy and resilience while employing a markedly reduced parameter count.

At present, bearing fault diagnosis models built by combining attention mechanisms, Transformer architectures, and convolutional networks have achieved certain improvements in feature extraction, interference resistance, and adaptation to small-sample scenarios. However, these methods still exhibit significant shortcomings in real-world complex industrial operating conditions, and their overall diagnostic performance struggles to meet practical engineering requirements. Most

models employ only a single time-domain or frequency-domain attention branch, relying solely on features from one domain to perform unidirectional filtering of features from the other domain. This prevents the mutual guidance and correction of time-domain and frequency-domain information. This mode of information interaction results in time-frequency heterogeneous features operating independently and information disconnection, making it impossible to fully extract and utilize subtle fault features arising from the coupling of temporal changes and frequency energy within vibration signals.

At the same time, traditional convolutional architectures generally use fixed-size convolutional kernels for feature extraction, making them unable to adaptively capture multiscale information based on the actual characteristics of bearing fault signals. Key features of varying granularity—whether fault impact pulses, periodic harmonic components, or low-frequency trend changes—are prone to information loss during computation.Currently, mainstream dual-stream networks also generally suffer from simplistic fusion strategies; most simply concatenate time-domain and time-frequency features directly in the final stage of the network, lacking cross-domain information exchange and preprocessing steps prior to fusion.Environmental noise and dynamic fluctuations in operating conditions within industrial settings generate a large amount of redundant features, which severely interfere with the fusion of effective features. This leads to a significant decline in both diagnostic accuracy and generalization performance when the model is applied to complex scenarios involving variable speeds, variable loads, and uneven sample distributions. Based on the practical issues of existing methods—such as insufficient time-frequency interaction, incomplete multiscale feature extraction, and the lack of pre-calibration in dual-stream feature fusion—this paper proposes the following:

(1) A bidirectional cross-attention module is designed for time-frequency domain interaction, which breaks free from the constraints of traditional single-branch attention mechanisms to enable bidirectional guided attention computation between time-domain and frequency-domain features. Through feature projection, cross-domain attention weight learning, and a gated fusion mechanism, time-domain features guide frequency-domain features in mining key information, while frequency-domain features conversely reinforce the fault representation of time-domain features, thereby effectively enhancing the cross-domain expressiveness of fault-sensitive features.

(2) A multi-branch, multi-scale one-dimensional convolutional module is constructed. Fine-grained, medium-grained, and coarse-grained features are extracted from vibration signals through three convolution kernels with sizes of 3, 15 and 31. These multi-branch features are fused by means of 1×1 convolution. The limitation that a single convolution kernel fails to cover fault features of diverse scales is overcome by this scheme, and fault information contained in different frequency components can be extracted more comprehensively.

(3) A dual-branch network architecture is established, comprising a one-dimensional time-domain stream and a two-dimensional time-frequency stream. The time-domain stream focuses on the temporal variation characteristics of the signal,while the time-frequency stream explores energy distribution characteristics in the frequency domain. With bidirectional cross-attention at its core, the architecture first performs cross-domain interactive enhancement of time-frequency features, then achieves dual-branch feature fusion through global concatenation. This allows time-frequency heterogeneous features to be mutually calibrated prior to fusion, further enhancing the complementarity of features across the time and frequency dimensions.

The organization of the subsequent sections of this paper is outlined as follows. Section 2 introduces related work on convolutional neural networks and attention mechanisms. Section 3 provides a detailed explanation of the proposed TFBCAF network architecture, covering the design concepts and specific implementation details of the coordinate attention module, multiscale one-dimensional convolution module, dual-stream feature extraction branches, and the fusion classification layer. Section 4 validates the model's effectiveness through multiple sets of comparative experiments, comparing fault detection accuracy across different attention mechanisms, feature fusion methods, and training strategies, while also analyzing the model's performance in identifying various fault types. Section 5 summarizes the research findings of this paper, highlighting the model's advantages and identifying areas for future improvement.

https://doi.org/10.37965/jdmd.2026.1840

2. Theoretical Background

2.1 Time-Frequency Domain Interactive Bidirectional Cross-Attention Mechanism

Most traditional attention mechanisms can only enhance features within a single domain or branch and are unable to achieve interaction and mutual guidance between features across different domains. Considering the practical needs of time-frequency dual-stream fusion, this paper designs a time-frequency domain interactive bidirectional cross-attention mechanism,This bidirectional attention module, spanning both the time domain and the time-frequency domain, undergoes three core stages—feature projection, bidirectional cross-attention computation, and gated fusion reconstruction—to enable mutual guidance, information exchange, and enhancement between time-domain and frequency-domain features. This design permits the two heterogeneous feature modalities to mutually calibrate essential information before being combined, thereby boosting the cross-domain representational capacity of fault signatures and addressing the practical demands of time–frequency dual-path rolling bearing diagnosis.

The core of the time-frequency domain interactive bidirectional cross-attention mechanism lies in the cross-domain application of multi-head self-attention. Concretely, both time-domain and frequency-domain features are first projected into an identical hidden dimensional space. Afterward, bidirectional attention is executed in two complementary manners: one where the time-domain signals serve as the keys and values while the frequency-domain signals act as the queries, and another where the frequency-domain representations form the key–value pairs and the time-domain representations function as the queries. A gating unit ultimately controls the blending proportion between the refined features and the original ones, thus achieving feature reconstruction and enhancement.

The core operation of multi-head self-attention is scaled dot-product attention. For queries $Q \in R^{B\times N_q\times D_q}$、$K \in R^{B\times N_k\times D_k}$ and $V \in R^{B\times N_k\times D_v}$ (where B is the batch size, N is the sequence length, and D is the feature dimension), the operation is as follows:

$$\text{Attention}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{1}$$

Multi-head self-attention applies linear transformations to queries, keys, and values, mapping them to multiple independent subspaces. After splitting them into multiple representation branches, it performs scaled dot-product attention operations in parallel, then concatenates and integrates the results from each group. This allows for the exploration of intrinsic relationships between cross-domain features across multiple dimensions. The specific computational expressions are as follows:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,...,\text{head}_h)W^O \tag{2}$$

$$\text{Head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{3}$$

Where W_i^Q,W_i^K,W_i^V,W^O is the learnable projection weight matrix, and h is the number of attention heads.

2.2 Feature Fusion

This hybrid architecture merges one-dimensional temporal signals with two-dimensional time–frequency representations, incorporating channel-wise squeeze-and-excitation recalibration along with a bidirectional cross-attention mechanism in the time–frequency domain; final integration is accomplished through concatenation at the feature level. Specifically, the SE module adaptively reweights the channel responses of both the time-domain and time–frequency branches, amplifying the contribution of components that are most sensitive to faults. The interactive bidirectional cross-attention component then facilitates mutual refinement between these heterogeneous modalities, enabling the exchange and alignment of critical information.

The two one-dimensional feature vectors are input into the time-frequency domain interactive bidirectional cross-attention module. Through their respective projection layers, they are mapped to the same hidden dimension to obtain projected features, and bidirectional cross-attention calculations are then performed to achieve mutual calibration of cross-domain features. Subsequently, the feature fusion layer concatenates and reconstructs the projected features with the

https://doi.org/10.37965/jdmd.2026.1840

attention-weighted features.

For the feature map $F_t \in R^{C_t \times L_t}$ extracted from the 1D time-domain branch, adaptive average pooling is applied to obtain the global time-domain feature vector $v_t \in R^{C_t \times 1}$, which is then compressed into a one-dimensional vector $v_t \in R^{C_t}$; for the feature map $F_f \in R^{C_f \times H_f \times W_f}$ extracted from the 2D time-frequency branch, adaptive average pooling is applied to obtain the global time-frequency feature vector $v_f \in R^{C_f \times 1}$, which is then compressed into a one-dimensional vector $v_f \in R^{C_f}$. The two one-dimensional feature vectors are concatenated along the channel dimension to obtain the fused feature vector:

$$v_{fusion} \in R^{C_t + C_f} \tag{4}$$

$$v_{fusion} = [v_t, v_f] \tag{5}$$

This fusion method, combined with the SE module and the bidirectional cross-attention enhancement mechanism, not only fully preserves the original information of time-domain and frequency-domain features but also achieves mutual calibration of cross-domain features and amplification of fault-sensitive information, enabling the two types of heterogeneous features to complement each other deeply. The combination of these three elements comprehensively characterizes the health status of bearings, addressing the issues of incomplete feature representation in single-modality approaches and misalignment of cross-domain feature information.The aggregated feature vector is subsequently passed to a dense classifier, enabling precise categorization of bearing defect types and significantly enhancing the model's diagnostic performance under demanding operating conditions.

3. Proposed Time-Frequency Dual-Stream Fusion Network

3.1 Overall Model Framework

The time-frequency bidirectional cross-attention fusion network adopts an end-to-end design approach. It comprises five core modules: one-dimensional time-domain feature extraction, two-dimensional time-frequency feature extraction, time-frequency domain interactive bidirectional cross-attention, time-frequency feature fusion, and a classifier. Data augmentation and regularization techniques are incorporated during training to enhance the model's robustness.

The model takes standardized one-dimensional rolling bearing vibration signals X_{raw} as input and ultimately outputs probability distributions corresponding to various faults. The specific forward inference process is as follows:

$$v_t = F_{1D}(X_{raw}) \tag{6}$$

$$v_f = F_{2D}(T_{STFT}(X_{raw})) \tag{7}$$

$$v_t^{enhance}, v_f^{enhance} = A_{T\text{-}F}(v_t, v_f) \tag{8}$$

$$v_{fusion} = C(v_t^{enhance}, v_f^{enhance}) \tag{9}$$

$$P = G(v_{fusion}) \tag{10}$$

where: $F_{1D}(\cdot)$ is the mapping function of the 1D time-domain feature extraction branch; $T_{STFT}(\cdot)$ is the short-time Fourier transform (STFT) time-frequency mapping function; $F_{2D}(\cdot)$ is the 2D time-frequency feature extraction branch mapping function; $A_{T\text{-}F}(\cdot)$ is the time-frequency domain interactive bidirectional cross-attention mapping function; $C(\cdot)$ is the feature fusion operation; $G(\cdot)$ is the classifier mapping function;

The specific workflow of TFBCAF is as follows:

(1) Input preprocessing: The raw vibration signal is normalized to eliminate differences in units and amplitude, resulting in a 1×1024-dimensional one-dimensional time-series input;

(2) For temporal feature extraction, the preprocessed vibration signal is routed into a 1D time-domain pathway, where a multi-scale 1D convolutional block employs kernels of three distinct sizes—3, 15, and 31—to capture fault-related patterns at varied granularities and then merges them. Subsequently, stacked 1D convolutional layers, integrated with batch normalisation and down-sampling operations, progressively learn deeper-level representations.These are fed into a one-dimensional SE module to enhance features from valid channels, and finally, global adaptive pooling is applied to generate a 1×128-dimensional time-domain global feature vectorv_t;

(3) Time-frequency Feature Extraction: The original signal undergoes an STFT transform; the transform results are

https://doi.org/10.37965/jdmd.2026.1840

log-amplitude transformed and then cropped or padded to form a 32×64-dimensional two-dimensional time-frequency spectrum, which enters the two-dimensional time-frequency branch. Multiple layers of two-dimensional convolutions, combined with batch normalization, pooling, and dropout operations, are used to extract spatial features from the time-frequency spectrum. These features are then fed into a two-dimensional SE module to enhance key channel features,and finally, global adaptive pooling is applied to generate a 1×128-dimensional time-frequency global feature vectorv_f ;

(4) For cross-domain attention-based enhancement, the global temporal descriptor and the global time–frequency representation are jointly passed into a bidirectional cross-attention module that interacts across both domains. Initially, the two vectors undergo a linear projection, followed by the ReLU activation and layer normalisation, mapping them into a common 64-dimensional latent space. Bidirectional interaction calculations are then executed in two modes: time-domain features serve as a guide for frequency-domain attention, and frequency-domain features steer time-domain attention. Finally, a sigmoidal gating coefficient adaptively determines the blending proportion of the enhanced and original representations, producing the refined temporal feature$v_t^{enhance}$ and$v_f^{enhance}$;

(5) Feature Fusion: The enhanced time-domain and frequency-domain features are concatenated along the channel dimension to form a 1×256-dimensional time-frequency fusion feature vectorv_{fusion} ;

(6) Fault Classification: The fused features are fed into a fully connected classifier. After passing through a 128-dimensional hidden layer, batch normalization, and the ReLU activation function, the output is regularized using Dropout. The output is a logarithmic probability distribution for each fault type, which is then mapped to the final probability distribution P via the Softmax function. The category corresponding to the maximum probability is selected to complete fault identification.

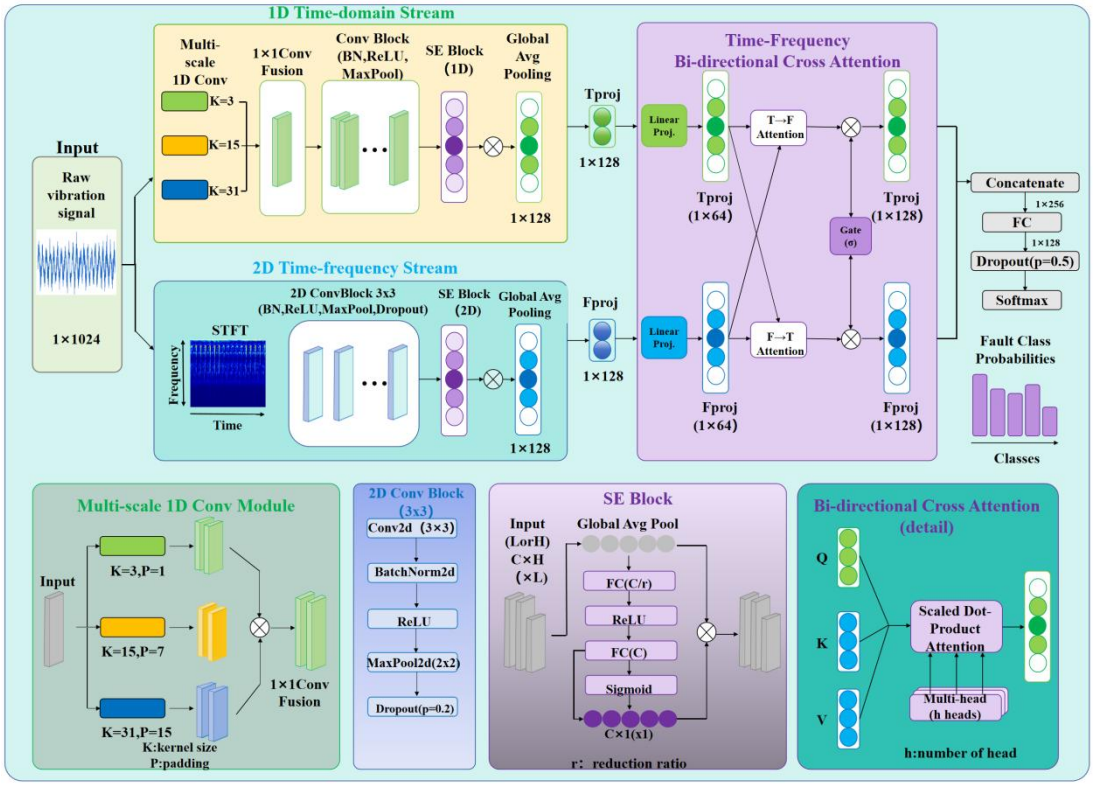


Figure 1: TFBCAFN Framework

3.2 TFBCAFN Module Design

The time-domain feature extraction unit is designed around the temporal characteristics of rolling bearing vibration signals, featuring a cascaded structure of multiscale convolutions combined with batch normalization, activation functions, and pooling. The unit takes the normalized one-dimensional vibration signal as input. It is configured in parallel with three different-sized convolutional kernels—3, 15, and 31—to cover fault features across different frequency ranges: small-sized

kernels capture high-frequency local impacts, medium-sized kernels extract mid-frequency modulation information, and large-sized kernels uncover low-frequency global trends. The outputs from all branches are fused via a 1×1 convolution and unified into a 48-channel feature map. Subsequently, the number of channels is gradually increased to 128 dimensions through two convolution operations. During this process, batch normalization and the ReLU activation function are applied to enhance feature expressiveness, while pooling operations with a stride of 4 are used to reduce the feature dimension while retaining key information. Finally, adaptive average pooling compresses the feature length to 1, yielding a time-domain global feature vectorv_t with dimensions $R^{128\times1}$. The core operation is multi-scale convolution fusion, expressed as:

$$F_{ms}=\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(\mathrm{Conv}_3(X),\mathrm{Conv}_{15}(X),\mathrm{Conv}_{31}(X)\right)\right) \tag{11}$$

3.2.1 Two-Dimensional Time-Frequency Feature Extraction Unit

The 2D time–frequency feature extraction pathway operates on spectrograms computed via the short-time Fourier transform (STFT) and follows a pipeline of time–frequency mapping, convolutional feature learning, and pooling-based down-sampling. Initially, the raw vibration signal is transformed by the STFT to obtain a joint time–frequency matrix. Logarithmic scaling is subsequently applied to amplify subtle fault-related signatures, after which the map is either clipped or zero-padded to a uniform size of 32 × 64.Specifically, the STFT adopts the Hann window. The window length is 256 points, the hop size is 64 points, and the FFT length is 256 points. After STFT calculation, we retain the first 128 frequency bins ranging from 0 Hz to the Nyquist frequency. We keep the first 32 frequency bins and truncate the time dimension to 64 time frames to construct the 32 × s64 spectrogram. This cropping strategy discards high-frequency components with weak fault-related information under variable speed conditions.

Three successive convolutional stages then expand the channel dimension from 1 to 128. Each convolution is followed by batch normalisation, a ReLU nonlinearity, and max-pooling with a stride of 2, collectively enabling deep discriminative feature extraction and progressive spatial reduction.a Dropout layer is also incorporated to mitigate the risk of overfitting. Finally, adaptive average pooling is used to compress the spatial dimensions to 1×1, yielding a 128-dimensional time-frequency global feature vectorv_f .

3.2.2 Time-Frequency Interaction Bidirectional Cross-Attention Module

Representing the key inventive contribution of the overall architecture, this block is specifically engineered to achieve bidirectional interplay and mutual enhancement between temporal and spectral representations. Its structural organization can be decomposed into three principal stages: projection of the features into a unified latent space, computation of bidirectional cross-attention, and gated reconstruction via adaptive fusion.

In our implementation, both the time-domain and time-frequency-domain branches output a single global descriptor after global adaptive average pooling. To satisfy the sequence-form input requirement of multi-head attention, each standalone global vector is formed into a feature sequence of length one. Consequently, the query, key and value sequences all have a sequence length of one. Under this configuration, multi-head attention cannot perform selection among multiple key feature elements, and the weights yielded by softmax normalization are fixed to one. Nevertheless, the learnable projection matrices within the multi-head architecture are still capable of cross-modal subspace transformation. The performance gains brought by this module stem from multi-head subspace mapping, concatenation-based reconstruction network and learnable gating mechanism, rather than dynamic weighting over multiple key feature elements.

(1) Feature Projection: Spatial Alignment and Nonlinear Enhancement. The original 128-dimensional time-domain feature vector and frequency-domain feature vector are fed into a network structure that incorporates a fusion linear transformation, layer normalization, and the ReLU activation function. Through this process, both types of features are uniformly mapped to a 64-dimensional hidden space, successfully achieving spatial alignment and nonlinear enhancement optimization of the different features.

(2) Bidirectional cross-attention computation: Once the temporal and spectral descriptors have been projected into a

shared latent space, their interaction is orchestrated through a two-way multi-head attention scheme. This arrangement permits each feature type to alternately fulfill the query, key, and value roles for the other, enabling reciprocal information transfer. As a consequence, the spectral representations become infused with time-varying dynamics extracted from the temporal dimension, while the temporal descriptors assimilate frequency distribution patterns derived from the spectral domain, thereby achieving mutual complementarity and enhancement between the two representational spaces.The tensor dimensions at this stage are as follows: batch size B, hidden dimension 64, number of attention heads 4, per-head dimension d_k=16. Both projected temporal and frequency-domain features have dimensions of $B \times 1 \times 64$. The softmax operation is performed along the feature sequence dimension.For the above formulas, the sequence lengths of both query and key equal one. Here the multi-head attention mainly serves as multi-head joint feature projection instead of standard multi-element cross-attention.

$$F_{att}=\text{MultiHeadAtt}(F_{proj},T_{proj},T_{proj}) \tag{12}$$

$$T_{att}=\text{MultiHeadAtt}(T_{proj},F_{proj},F_{proj}) \tag{13}$$

(3) Gated Fusion and Feature Reconstruction: The model concatenates the projected original features with the enhanced features optimized by the attention mechanism, then restores the concatenated features to their original dimension. Through learnable gating parameters combined with a Sigmoid function to constrain weight values, the model autonomously adjusts the fusion ratio between the original and enhanced features. This approach enhances the stability and classification discrimination capabilities of the model's features while preserving as much effective original information as possible.The global time-domain and frequency-domain features are separately fed into projection networks and mapped to identical hidden dimensions. Bidirectional multi-head feature transformation is then executed, where two modal features alternately act as query and key-value pairs. The transformed outputs are concatenated with original projected features and sent to the reconstruction network to recover the original feature dimension. Finally, learnable gating coefficients adaptively fuse raw and enhanced features to produce updated global time-domain and frequency-domain features.

$$g_t=\sigma(\alpha_t),\quad g_f=\sigma(\alpha_f) \tag{14}$$

$$T_{out}=g_t \cdot T_{enhance}+(1-g_t)\cdot T \tag{15}$$

$$F_{out}=g_f \cdot F_{enhance}+(1-g_f)\cdot F \tag{16}$$

3.2.3 Feature Fusion and Classifier

During the fusion step, channel-wise concatenation of the enhanced temporal vector and the time–frequency representation yields a composite feature vector of size 256×1, thereby fully retaining the unique characteristics embedded within the two original modalities. For the classification component, a two-tier dense architecture is built: the initial 256-dimensional fused output is first projected into a 128-dimensional intermediate representation, after which its dimensionality is tailored to the exact number of fault categories. In the forward pass, batch normalisation, ReLU nonlinearity, and a dropout layer operating at a rate of 0.5 are consistently applied. This approach not only performs nonlinear feature transformations but also effectively controls model overfitting. The predicted score vector output by the network is processed through the Softmax function to calculate the probability results corresponding to each fault category, thereby completing the fault identification task. The corresponding computational expressions are as follows:

$$P[i,j]=\frac{e^{y_{pred}[i,j]}}{\sum_{k=1}^{N} e^{y_{pred}[i,k]}} \tag{17}$$

3.3 Optimization Process

3.3.1 Data Preprocessing and Augmentation

During the data preprocessing stage, Z-score normalization was applied to eliminate dimensional differences in the raw vibration signals and ensure numerical stability during model training. The formula is as follows:

$$\hat{X}_{raw}=\frac{X_{raw}-\mu(X_{raw})}{\sigma(X_{raw})+10^{-6}} \tag{18}$$

https://doi.org/10.37965/jdmd.2026.1840

where: μ(•) and σ(•) represent the mean and standard deviation of the original signal, respectively; 10^{-6} is used to avoid the exceptional case of a zero denominator.For data augmentation during model training, random amplitude scaling is adopted. For each training sample, the raw vibration amplitude is multiplied by a random scaling factor sampled from [0.9, 1.1] with a probability of 50%. No other data augmentation operations are utilized in this work.

3.3.2 Loss Functions and Optimization Settings

For the model classification task, we adopt a label-smoothed cross-entropy loss function. By smoothing the label distribution, this function mitigates the model's overconfidence in the training data and reduces the risk of overfitting. Its expression is:

$$L_{LSCE}=-\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{N}\left[(1-\epsilon)\cdot y_{ik}+\frac{\epsilon}{N-1}\cdot(1-y_{ik})\right]\cdot \log(p_{ik}) \tag{19}$$

where: B is the number of samples in a batch; N is the number of fault classes;$\epsilon = 0.1$ is the label smoothing coefficient; y_{ik} is the raw hard label of the i-th sample; p_{ik} is the probability predicted by the model that the i-th sample belongs to class k. This loss function is compatible with different versions of the PyTorch framework; when label smoothing is natively supported, it is called directly; otherwise, it is implemented via a custom solution.

The Adam algorithm was selected for model optimization. Leveraging the complementary benefits of momentum and adaptive step-size strategies, the optimizer accomplishes efficient weight adjustments over the entire training duration. In the experimental configuration, the starting learning rate η was fixed at 0.001, while the weight decay parameter was assigned a value of 10^{-4} to restrain excessive parameter expansion. A Cosine Annealing LR schedule was adopted throughout training to avert convergence toward suboptimal local minima in the later epochs; under this scheme, the learning rate follows a periodic decay governed by the training epoch index according to a cosine function, as formalized in the following expression:

$$\eta_t=\eta_{min}+\frac{1}{2}(\eta_{max}-\eta_{min})\cdot\left(1+\cos\left(\frac{t}{T_{max}}\cdot\pi\right)\right) \tag{20}$$

where: $\eta_{max} = 0.001$and $\eta_{min} = 10^{-6}$are the maximum and minimum values of the learning rate, respectively; t is the current training iteration; and `$T_{max} = 150$ is the maximum number of training iterations.

4. Experimental Validation

4.1 Experimental Data

A simulated experimental setup was built to collect data on bearing defects across various operating conditions [25], as illustrated in Figure 2. The dataset specifics and model training details are presented in Table 1. The essential elements of this configuration comprise an electric motor, a turntable, a bearing holder, a gearbox, and a brake system. The acceleration sensor was mounted directly on the bearing housing, whereas the acoustic sensors were located around 10 centimeters away from it. The raw long-duration vibration signals are acquired under continuously varying rotational speed (800-1500 rpm), instead of assembling multiple independent constant-speed signal segments. Each fault category yields 500 non-repeated signal segments segmented from the continuous recording. A representative speed-time curve sample is supplemented in Figure 2 to demonstrate the continuous speed-fluctuation characteristic. A 5:5 stratified random split is further adopted to construct training set and test set. Signal fragments in training and test sets have no sample-level duplication, while their rotational-speed value ranges still overlap within 800-1500 rpm. No cross-validation is carried out for rotational speeds fully outside the sampled interval.Vibration acceleration sensors are affixed to the base and bearing seat surfaces to enable data gathering, subsequently recorded using the LMS data acquisition system. Figure 3 shows the raw vibration signals and

time–frequency spectrograms for four fault types: Normal condition (NC); Inner ring failure(IF); Outer ring failure(OF); Rolling element failure(RF).The numeric suffixes in labels like IF0.2/0.4/0.6 represent the defect size of bearing faults, with the unit of millimeters (mm). In this classification task, samples with different defect sizes at the same fault location are merged into one unified fault category. Specifically, IF0.2, IF0.4 and IF0.6 are all grouped into the inner-race fault category IF; similarly, samples of different defect sizes for outer-race and rolling-element faults are also merged respectively. Consequently, all samples are finally divided into four categories: normal condition (NC), inner-race fault (IF), outer-race fault (OF), and rolling-element fault (RF). This four-category setting is consistent with the confusion matrix and all subsequent experimental results.

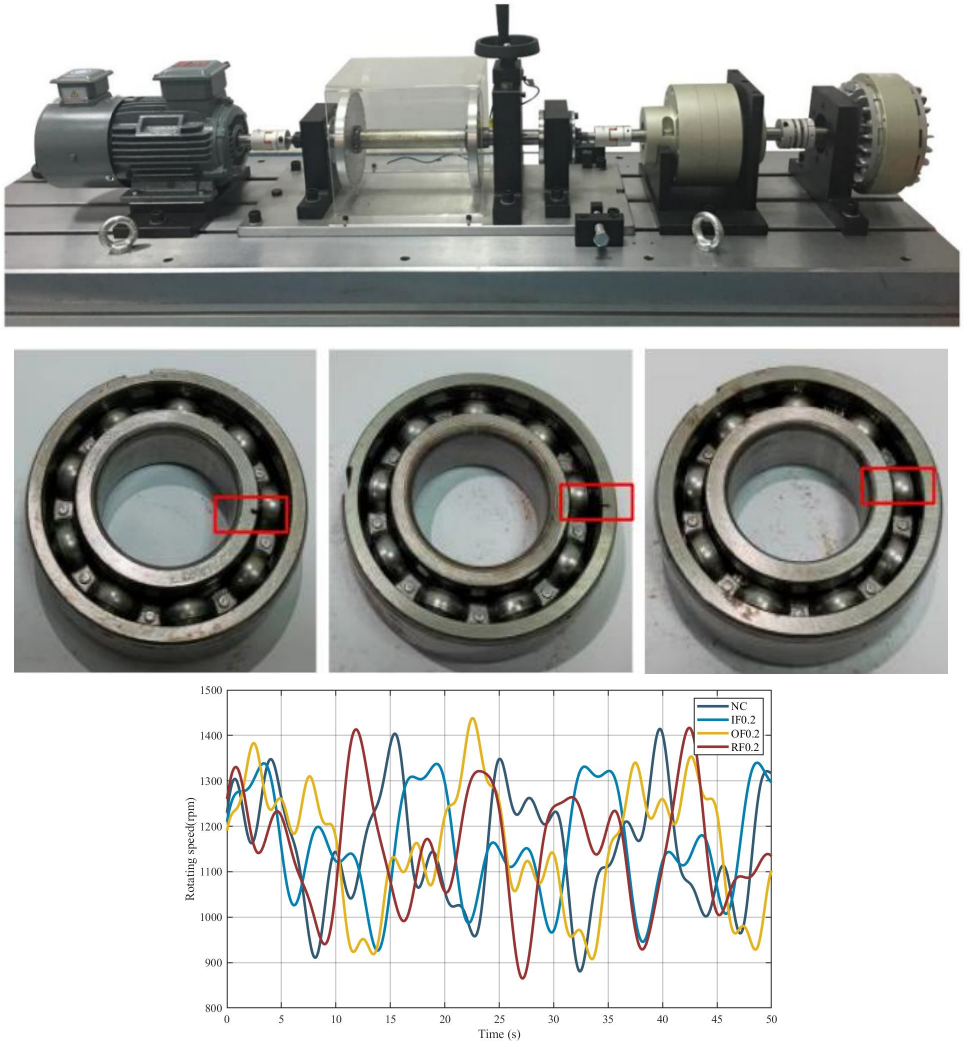


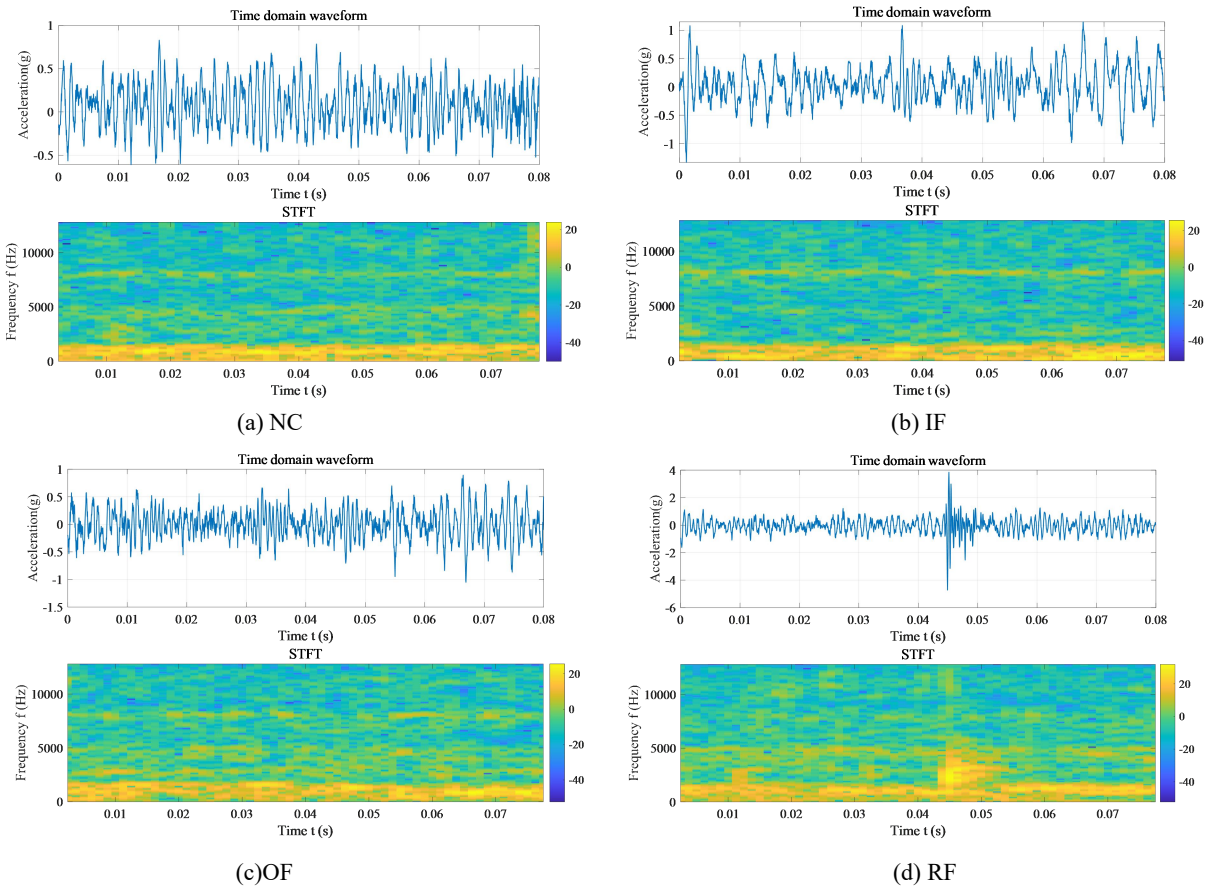
Figure 2: Experimental Platform and Faulty Bearing and Representative Speed-Time Curve of the Experimental Platform

Table 1: Dataset and Model Training Information

	Experiment1	Experiment2
Hyper parameter：γ β λ	1, 0.18, 0.000065	1, 0.166, 0.00005
Learning rate	0.001	0.001
The quantity of training iteration	150	150

https://doi.org/10.37965/jdmd.2026.1840

cycles		
The quantity of each fault.	500	500
Failure mode	NC,IF0.2/0.4/0.6,OF0.2/0.4/0.6, RF0.2/0.4/0.6	NC,IF0.2/0.4/0.6,OF0.2/0.4/0.6, RF0.2/0.4/0.6
Load conditions	0	0
Speed conditions	800-1500rpm	800-1500rpm
Sampling frequency	25.6kHz	25.6kHz
Bearing type	6205	NU205EM



(a) NC (b) IF

(c)OF (d) RF

Figure 3 : Time–frequency representations of raw vibration signals under four fault conditions

4.2.1 Experiment 1: Deep Groove Ball Bearing 6205

In all trials of Experiment 1, evaluations were performed under identical environmental settings. The study encompassed a comparative assessment of the classification accuracy achieved by six distinct bearing fault diagnosis models, aiming to gauge the practical capability of the TFBCAF model in analogous recognition scenarios. Five benchmark models that are widely adopted in the field were chosen as baselines to confirm the superiority of the approach introduced in this work. All compared models including the proposed hybrid network, all ablation variants are trained under unified hyper-parameter settings to ensure fair comparison. The Adam optimizer is utilized. The initial learning rate is set to 0.001, and the cosine annealing learning rate scheduler is adopted with the minimum learning rate of 1×10^{-6}. The batch size is 64, and the maximum training epoch is 150. Cross-entropy loss with label smoothing of 0.1 serves as the

https://doi.org/10.37965/jdmd.2026.1840

loss function. During training, the model checkpoint achieving the highest test accuracy is preserved for final performance evaluation.The architectural details of these comparative models are summarized below:

Method 1: 1D-CNN: A one-dimensional convolutional neural network, a classic method for directly processing vibration time-series signals.

Method 2: 2D-CNN: A method that converts one-dimensional signals into two-dimensional time-frequency plots and then uses two-dimensional convolution for feature extraction.

Method 3: CNN-BiLSTM: A hybrid model combining a convolutional neural network with a bidirectional long-short-term memory (BiLSTM) network, which balances local features and temporal dependencies.

Method 4: CNN-Transformer: A model that uses a CNN to extract local features and a Transformer to capture global dependencies.

Method 5: CNN-BiTCN-CrossAttn: A feature fusion model that combines a bidirectional temporal convolutional network with a cross-attention mechanism.

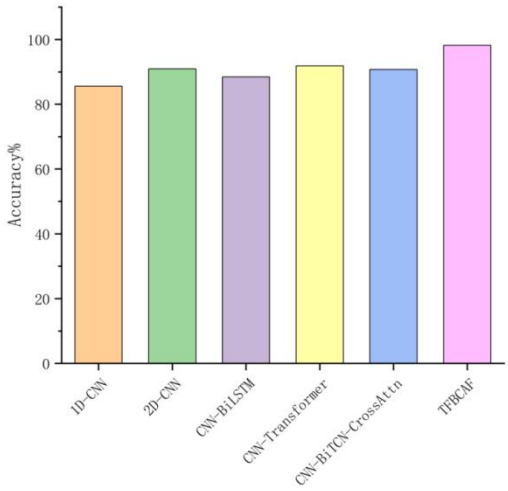


Figure 4 : Histogram of accuracy distributions for the six methods

Figure 4 illustrates the bar chart summarized from accuracy results of six comparative bearing fault identification experiments. Figure 5 shows violin plots of accuracy distributions for the six methods. By comparing the measured performance of TFBCAF with the other five baseline models based on this data, it is evident that the model proposed in this paper outperforms all comparison algorithms in overall recognition performance. The entire chart records the test accuracy values for all models after 150 iterations; with the exception of our self-developed model, the final accuracies of the remaining comparison algorithms all show a significant gap. Networks built solely on a single modality, such as 1D-CNN and 2D-CNN, achieved final accuracies of 85.86% and 90.88%, respectively.while the remaining models—CNN-BiLSTM, CNN-Transformer, and CNN-BiTCN-CrossAttn, which incorporate temporal structures or attention modules—achieved test accuracies of 88.44%, 91.84%, and 90.72%, respectively; however, their performance still fell short of the TFBCAF model across all metrics. The trends in the data reveal that the 1D-CNN ranks last in accuracy among all algorithms; as it relies solely on one-dimensional temporal data, the 1D-CNN is limited by the lack of frequency-domain information and thus struggles to achieve ideal recognition performance. , equipped with a cross-attention structure, achieves a significant performance leap thanks to its multi-feature fusion design, and the results further demonstrate the practical effectiveness of cross-modal information interaction in improving fault recognition accuracy. Therefore, the TFBCAF can simultaneously integrate both one-dimensional time-series and two-dimensional time-frequency features. By utilizing cross-attention to match and complement features from different modalities, its approach of blending information from both dimensions fully captures the latent correlations within the signal. The information pathways established by cross-modal attention further expand the range of features the model can extract, and this comprehensive design enabled the model to achieve a recognition accuracy of 98.2% in the experiments.

https://doi.org/10.37965/jdmd.2026.1840

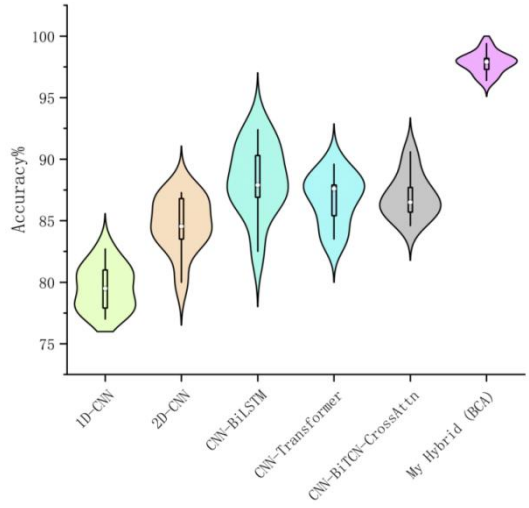


Figure 5: Violin Plot of Accuracy Distributions for Six Methods

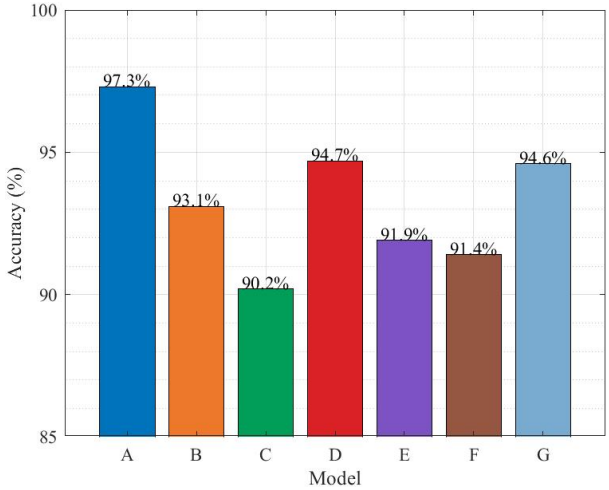


Figure 6: Ablation Experiment Diagram

The model structure is disassembled stepwise via ablation experiments to analyze the influence of each core component on the overall performance of TFBCAF. The test accuracy data corresponding to different structural variants are consolidated in the respective charts. The same training parameters were maintained throughout the entire experiment to ensure that the results of testing the individual modules' effects are meaningful. Prototype Model A, which includes all components in their full configuration, achieved the highest recognition accuracy across all categories in this set of experiments.

To verify the function of each module, multiple variant models are constructed for comparative analysis:

Model B (w/o BCA): Retained only the self-attention mechanism and removed the bidirectional cross-attention (BCA) module; accuracy dropped to 93.1%;

Model C (w/o Multi-Scale Conv): Removed the multi-scale convolutional module, resulting in an accuracy of 90.2%;

Model D (Single Stream, Time Only): Used only one-dimensional time-series signals as input, achieving an accuracy of 94.7%;

Model E (Single Stream, Time-Frequency Only): Uses only two-dimensional time-frequency representations as input, with an accuracy of 91.9%;

Model F (Unidirectional Attention): Replacing cross-attention with unidirectional attention resulted in an accuracy of 91.4%.

Model G (Direct Concatenation Late Fusion): Retained dual feature-extraction branches together with SE blocks, and replaced the interaction module with direct concatenation of two pooled descriptors; accuracy reaches 94.56%;

The comparison results show that removing any of the key modules leads to a varying degree of decline in the model's

diagnostic accuracy, with the performance loss resulting from removing the multi-scale convolution or cross-attention mechanism being the most significant. This validates the necessity and effectiveness of multi-scale convolution feature extraction, bidirectional cross-attention, and the 1D+2D bimodal fusion strategy in enhancing the model's fault diagnosis performance.

λ stands for the L2 regularization coefficient, and β refers to the gate initialization value. β = 0.18 and λ = 0.000065 correspond to bright regional colors and the best overall performance;When β is set too low, the gating mechanism favors the original features, suppressing the effect of attention; the entire figure appears cooler in color, and accuracy is lower; when β is set too high, the model over-relies on attention features, losing valuable original information, resulting in a slight decrease in accuracy; if λ is too small, significant overfitting occurs—training performance is good but test performance declines; if λ is too large, the regularization constraint is too strong, limiting the model's expressive power and causing overall accuracy to drop accordingly.

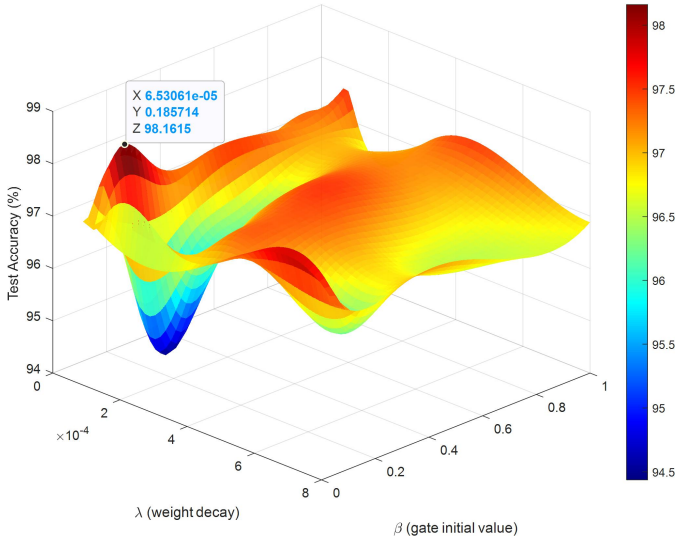


Figure 7: Model Parameter Ablation Experiment

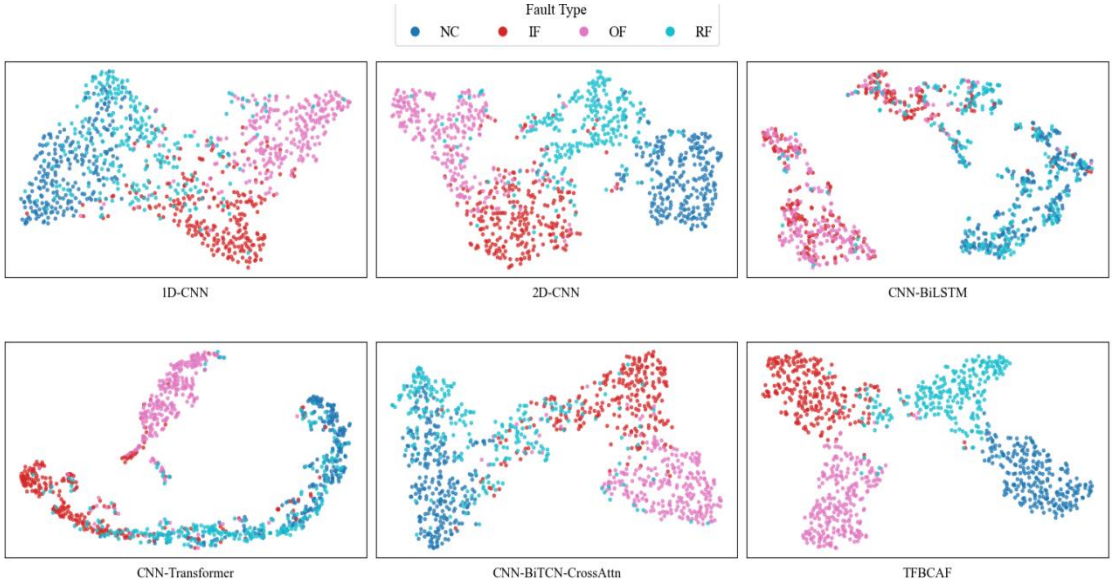


Figure 8: T-SNE Visualization

T-SNE is commonly used for dimensionality reduction and visualization of high-dimensional data. In this test, high-dimensional features are projected onto a two-dimensional plane as show in Figure 8, allowing the underlying clustering patterns and spatial relationships among samples to be intuitively visualized in the distribution plot. The feature distributions of all comparison models after dimension reduction are summarized in the corresponding figures. The feature

https://doi.org/10.37965/jdmd.2026.1840

points generated by 1D-CNN and 2D-CNN are scattered, with samples corresponding to different faults extensively overlapping. These two networks, which rely solely on single-modal convolutions, struggle to accurately classify various fault samples.The features generated by CNN-BiLSTM, CNN-Transformer, and CNN-BiTCN-CrossAttn show improved clustering compactness, with samples of the same category clustering more effectively. However, the boundaries between some categories remain blurred, and a small number of samples are still intermixed. Among these, the CNN-BiTCN-CrossAttn model achieves relatively better clustering results, though a small number of cross-category sample mix-ups still occur.

In contrast, the feature clustering performance of the TFBCAF model proposed in this paper is the most ideal. Sample points from different classes achieve high intra-class clustering and clear inter-class separation in the two-dimensional space, with very little cross-class mixing. This indicates that the fusion strategy combining 1D+2D hybrid features with a bidirectional cross-attention mechanism can effectively extract highly discriminative fault features, achieving superior feature alignment and class discrimination. Comparative experiments confirm that models with insufficient feature discriminative power and severe inter-class confusion suffer from significantly degraded fault diagnosis performance, whereas the highly discriminative features extracted by our method are the key reason for its excellent diagnostic accuracy.

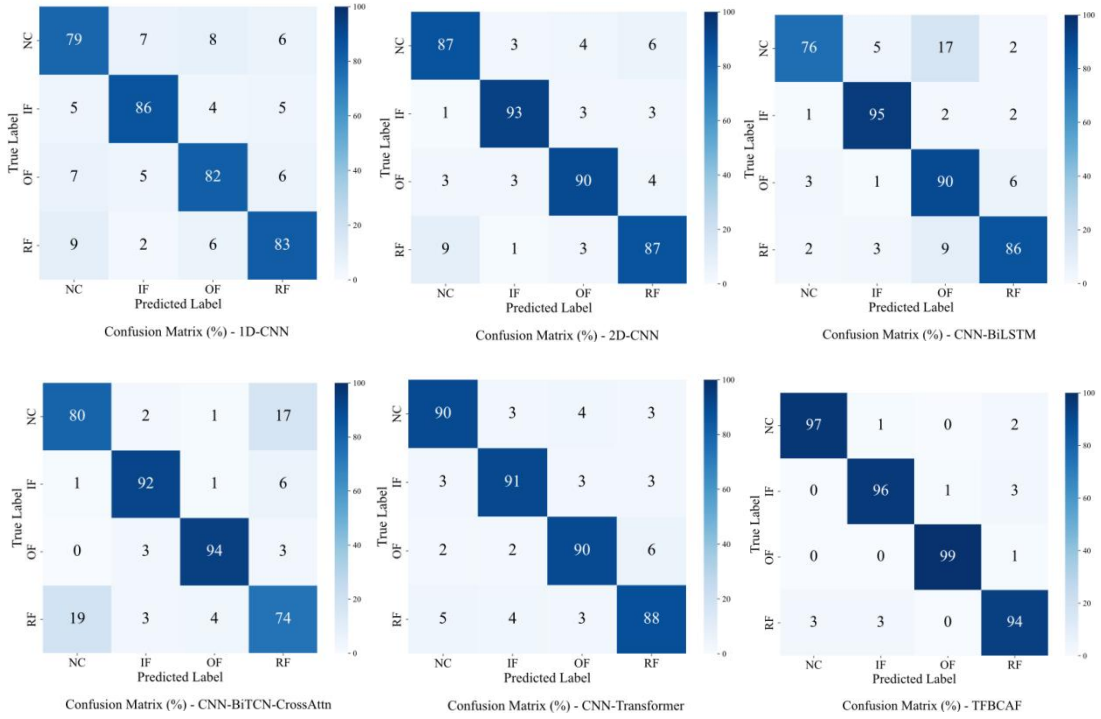


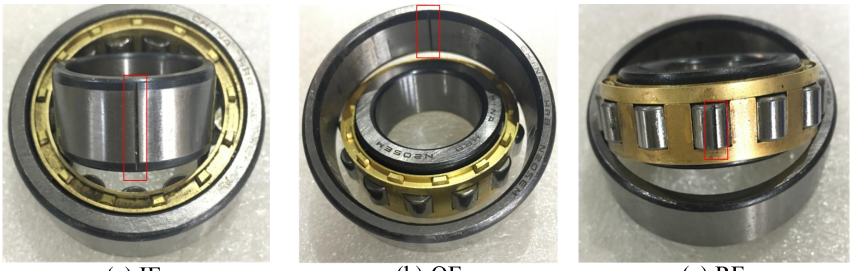
Figure 9: Confusion matrix comparing the results of Experiment 1's various comparison tests with those of the TFBCAF classifier

Figure 9 presents the confusion matrix results for six bearing fault diagnosis models, clearly illustrating each model's classification performance and misclassification patterns across different fault categories. The five baseline models—1D-CNN, 2D-CNN, CNN-BiLSTM [12], CNN-BiTCN-CrossAttn [31], and CNN-Transformer [32] —all exhibit misclassification issues to varying degrees: 1D-CNN has low classification accuracy for NC and RF, and a significant number of RF samples are misclassified as NC;CNN-BiTCN-CrossAttn exhibits significant misclassification in the RF class, with a large number of samples incorrectly classified as NC; although CNN-Transformer shows improved overall accuracy, it still exhibits a small amount of cross-class confusion. In contrast, the TFBCAF model has the highest diagonal values in its confusion matrix: 97% for the NC class, 96% for the IF class, 99% for the OF class, and 94% for the RF class,with very few misclassified samples—only minor confusion between NC and RF, and between IF and RF. Its overall classification performance is significantly superior to that of other baseline models. In particular, it achieved nearly zero misclassification for OF faults. This demonstrates that the architecture incorporating bidirectional time-frequency cross-attention fusion can

more effectively capture the feature differences among various faults, thereby enhancing the model's classification robustness and generalization ability. TFBCAF's classification performance in this set of experiments was markedly superior to that of the other approaches; classification errors occurred only in isolated instances across the entire dataset, and widespread cross-class confusion was largely avoided.

4.2.2 Experiment 2: Cylindrical Roller Bearing NU205EM

In this experiment, the NU205EM cylindrical roller bearing was selected as the test object. The runtime environment, training parameters, and control model configurations from the previous experiment were retained throughout. By comparing the test accuracy of different algorithms across multiple datasets, we evaluated the TFBCAF model's adaptability and operational stability in various bearing fault recognition scenarios. Figure 10 shows the bearing diagrams of the NU205EM cylindrical roller bearing under Inner ring failure(IF); Outer ring failure(OF); Rolling element failure(RF).



(a) IF (b) OF (c) RF

Figure 10: NU205EM cylindrical roller bearings

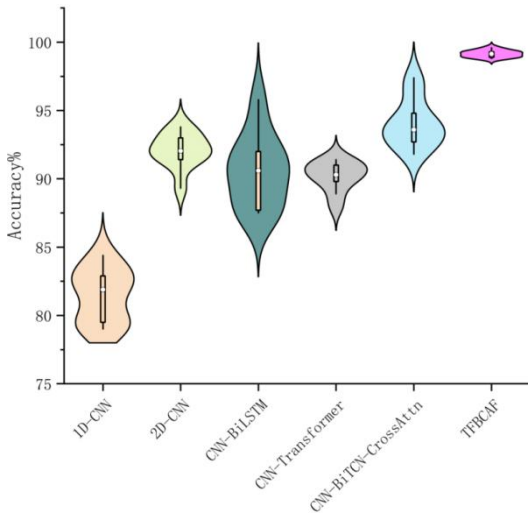


Figure 11: Violin plot of accuracy distributions for the six methods

Figure 11 displayed the violin plot of accuracy for the six methods. Analysis of the experimental results after adjusting the sample length shows that, when the sample length was set to 2048, the diagnostic accuracy of all models for the cylindrical roller bearing dataset increased.The 1024-point sample length in Section 3.1 corresponds to the baseline bearing dataset. In contrast, the variable-speed dataset in this subsection adopts 2048-point samples. We adopt different sample lengths to distinguish the two experimental setups and verify the generalization capability of the proposed network across different signal acquisition configurations. Longer signal segments capture more complete fault characteristic cycles under speed fluctuations, facilitating the extraction of robust fault features for variable-speed diagnosis.

However, significant differences in recognition capabilities still exist among the models. The diagnostic performance of 1D-CNN and 2D-CNN models, built on a single modality, remained consistently weak. The three comparison

https://doi.org/10.37965/jdmd.2026.1840

models—which incorporated temporal modules and basic attention structures—showed improved diagnostic accuracy. However, their feature fusion methods were relatively limited, failing to deeply mine effective information from both the time and frequency domains, thus hindering further improvements in diagnostic performance. The TFBCAF model utilizes a time-frequency dual-branch architecture combined with bidirectional cross-attention to optimize feature interaction. Its diagnostic accuracy outperforms all comparison algorithms, and the model's accuracy data exhibits greater convergence. It maintains stable and reliable diagnostic performance across various bearing failure recognition scenarios for different bearing models.

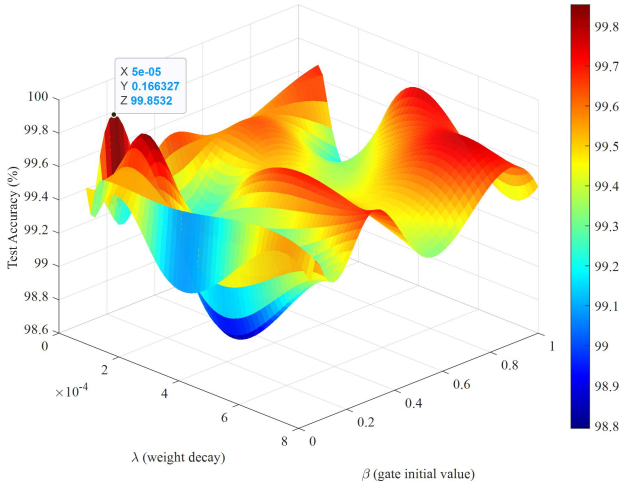


Figure 12: Model Parameter Ablation Experiment

Figure 12 showed the model parameter ablation experiment results intuitively illustrates the fluctuating effects of two parameters—the L2 regularization coefficient λ and the gating initialization value β—on the model's fault diagnosis accuracy. The experiments determined that the optimal parameter combination is β = 0.166 and λ = 5e-05; under these conditions, the model's diagnostic accuracy reaches 99.85%. If the value of β is too small, the weights of the gating module will be biased toward the original input features, making it difficult for the bidirectional cross-attention mechanism to optimize the features. If the value is too large, the model will over-rely on the attention-optimized features, discarding the inherent effective features of the vibration signal, which ultimately reduces diagnostic accuracy. A low value of λ fails to provide sufficient regularization constraints for the network parameters, making overfitting likely during the training phase; conversely, a high value of λ overly constrains the network's learnable parameters, weakening the model's ability to autonomously extract fault features. The experimental results also confirmed the optimal hyperparameter values for this dataset.



Figure 13: T-SNE Visualization

https://doi.org/10.37965/jdmd.2026.1840

Using T-SNE dimension-reduction visualization results, we can intuitively compare the clustering performance of different models after fault feature extraction in Figure 13. The feature points extracted by 1D-CNN and 2D-CNN are sparsely distributed; different fault samples intermingle, resulting in poor distinguishability between categories and making it difficult to achieve precise fault classification. CNN-BiLSTM, CNN-Transformer, and CNN-BiTCN-CrossAttn improved the clustering of samples within the same category, but some samples from different fault categories still intermingled and overlapped, indicating a lack of differentiation in feature representation. The TFBCAF model outperformed the other comparison models in terms of visualization; samples from various fault categories clustered well in the two-dimensional projection space, with clear category boundaries and very few instances of sample confusion. Based on the visualization results, it can be concluded that the time-frequency dual-branch structure, combined with the bidirectional cross-attention module, can extract fault features with higher discriminative power and maintain stable and efficient feature extraction capabilities when adapted to different bearing dataset models.

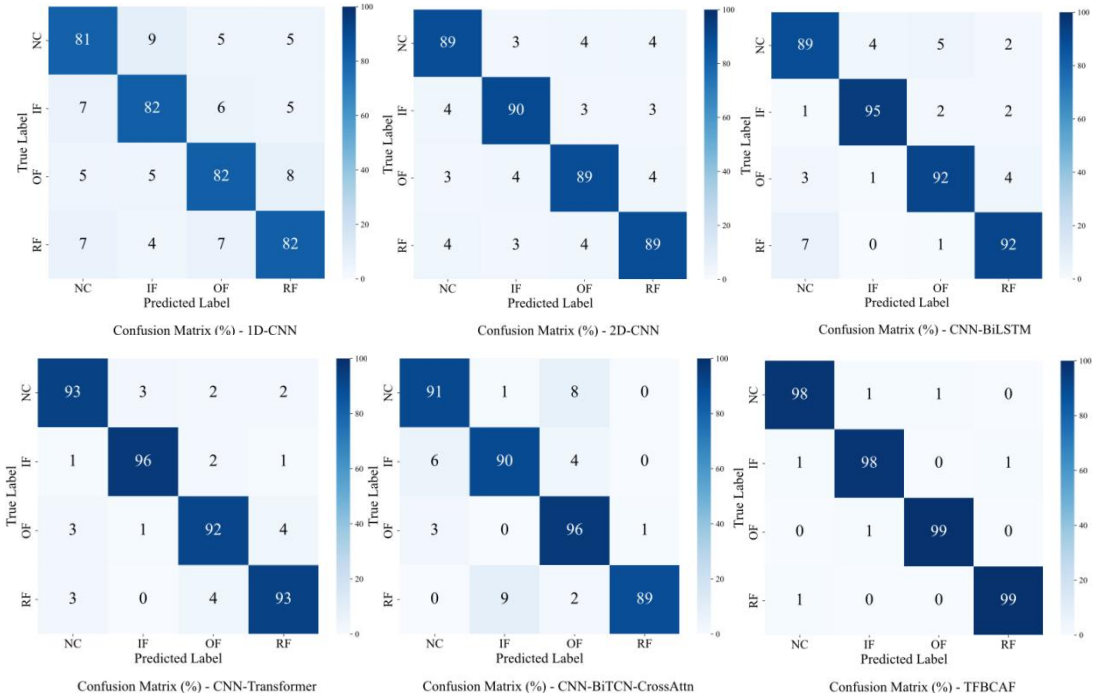


Figure 14: Confusion matrix comparing the classification results of Experiment 2's comparison models with the TFBCAF model

The confusion matrices for each model group intuitively reflect the classification results and misclassification patterns for the four types of bearing samples, as shown in Figure 14. Several models exhibited sample misclassification issues, with weak discriminative boundaries between fault categories.The diagonal values in the TFBCAF model's confusion matrix are generally high: The confusion matrix shows that the classification accuracy of inner race fault IF reaches 98%, while outer race fault OF and rolling element fault RF both achieve 99%.Only a small number of samples exhibited classification errors in the experiments, and overall, the model can accurately classify various types of bearing failures.

The experimental results demonstrate that, compared to traditional single-modality networks and conventional time-frequency hybrid networks, the TFBCAF model achieves superior diagnostic performance after being adapted to new bearing test data. This architecture employs multiscale convolutional kernels to capture temporal signal characteristics across varied resolutions, applies the STFT to inspect the spectral content of the vibration data, and integrates bidirectional cross-attention to forge an interactive pathway between the heterogeneous temporal and spectral representations, thereby overcoming the limitation of insufficient feature acquisition inherent in single-modality signals. By adaptively reweighting features through squeeze-and-excitation channel attention and regulating the training dynamics via regularization coefficients, the model accommodates fault recordings gathered from different bearing configurations. It preserves steady classification accuracy even when encountering previously unseen fault instances, rendering it suitable for practical industrial bearing condition monitoring tasks and exhibiting strong generalization capability.

5. Conclusion

To tackle the engineering difficulties of insufficient bearing fault signature depiction and limited data exploitation under varying speed conditions, this study introduces the TFBCAFN diagnostic framework built upon time–frequency feature integration. A bidirectional cross-attention unit is employed to form a mutual information exchange channel between temporal and spectral representations, thereby ameliorating the inherently weak expression of defect-related patterns in vibration measurements. The architecture adopts a parallel two-stream feature extraction design, where each stream is coupled with a dimension-matched squeeze-and-excitation channel attention block to refine and recalibrate the learned representations. By reallocating channel-wise importance scores, superfluous and uninformative components are suppressed, while discriminative attributes essential for fault recognition are amplified. Before the aggregated representation enters the classifier, the bidirectional cross-attention mechanism performs adaptive weighting of the bimodal streams, allowing time-domain and frequency-domain descriptors to mutually guide and supplement one another. This compensates for the informational gaps of unimodal data and accomplishes a deep-level fusion of heterogeneous time–frequency characteristics. Experimental findings reveal that, relative to conventional single-domain diagnostic approaches, the proposed model achieves favorable fault classification performance for non-stationary vibration signals with continuously fluctuating rotational speed within the sampled 800-1500 rpm range.Nevertheless, opportunities for further improvement persist in complex industrial environments involving dynamic load variations and intense noise contamination. Future research will concentrate on enhancing the adaptive regulation strategy of the time–frequency cross-attention mechanism under intricate and variable operating scenarios, as well as on devising superior fusion methodologies for multi-source modalities, such as vibration and acoustic signals.

Acknowledgement:

The research was supported by the Natural Science Foundation of Shandong Province (ZR2024ME236).

Declarations:

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported herein.

References:

[1] Zhang, Y., Shen, C., Ye, X., et al. A Dual-Component Elastic Adaptation Network for Rotating Machinery Incremental Fault Diagnosis under Variable Operating Conditions[J]. Journal of Dynamics, Monitoring and Diagnostics, 2026, 5(1), 49–63.

[2] Xin Y, Li S M, Wang J R. A New Fault Feature Extraction Method for Non-Stationary Signal Based on Advanced Synchrosqueezing Transform [J]. Journal of Vibration Engineering & Technologies, 2019, 7: 291–299.

[3] Yan S, Shao H, Wang X, et al. Compact-sparse prototype calibration network for few-shot continual fault diagnosis of rotating machinery[J]. Engineering Applications of Artificial Intelligence, 2025, 136, 109010.

[4] Wang J, Zhang X, Zhang Z, et al. Attention guided multi-wavelet adversarial network for cross domain fault diagnosis[J]. Knowledge-based systems, 2024, 284, 111285.

[5] Shao H, Ming Y, Liu Y, et al. Small sample gearbox fault diagnosis based on improved deep forest in noisy environments[J]. Nondestructive Testing and Evaluation, 2025, 40(8), 3935-3956.

[6] Hu M, Zhang X, Zhou T, et al. Noise-Vibration Monitoring and Analysis of Aero-Engine Blade Fracture Faults[J]. Journal of Dynamics, Monitoring and Diagnostics, 2026, 5(1), 86–95.

[7] Wu P, Yu G, Yu Q, et al. An adaptive few-shot fault diagnosis method based on virtual samples generated by fault characteristics of rotating machines[J]. Engineering Applications of Artificial Intelligence, 2024,136(PB): 109017-109017.

[8] Jiang X, Jun W, Shen C, et al. An adaptive and efficient variational mode decomposition and its application for bearing fault diagnosis[J]. Structural Health Monitoring, 2021, 20(5): 2708-2725.

[9] Mutra R, Reddy D, Rao K,Crack fault diagnosis in rotor bearing system by transient and study state time domain analysis[J]. Measurement, 2025, 241, 115667.

[10] Yang J, Yan L, Xu J, et al. Small-shot mechanical fault diagnosis based on multiscale expandable convolution and DMWT-Mamba

[J]. Noise and Vibration Control, 2026, 46(1):142–148+246.

[11] Wang Z, Fan J, Guo Y, et al. Research on Fault Feature Extraction for Planetary Bearings Using a Novel Combination of Operator-Based Morphological Filtering and Stochastic Resonance [J]. Journal of Sound and Vibration, 2025, 607: 119037.

[12] Tian H, Fan H, Feng M, et al. Fault diagnosis of rolling bearings based on an HPSO-optimized CNN-LSTM neural network [J]. Sensors, 2023, 23(14): 6508.

[13] Liu X, Xia L, Shi J, et al. A fault diagnosis method for rolling bearings based on an improved recurrence plot and a convolutional neural network[J]. IEEE Sensors Journal, 2023, 23(10): 10767–10775.

[14] Lu J, Wang K, Chen C, et al. A deep learning method for rolling bearing fault diagnosis based on the attention mechanism and Graham angle field [J]. Sensors, 2023, 23(12): 5487.

[15] Zhang Z, Jiang Q, Zhan Y, et al. Fault Classification Method for Small-Shot Rolling Bearings Based on the VAE-GAN Data Augmentation Algorithm [J]. Atomic Energy Science and Technology, 2023, 57(S1):228-237.

[16] Mohiuddin M, Islam M, Islam S, et al. Intelligent Fault Diagnosis of Rolling Element Bearings Based on Modified AlexNet [J]. Sensors, 2023, 23(18): 7764.

[17] Kim Y, Kim K. Time-frequency multi-domain 1D convolutional neural network with channel-spatial attention for noise-robust bearing fault diagnosis [J]. Sensors, 2023, 23(23): 9311.

[18] Zhao Z, Zhang R, Liu K, et al. A Small-Sample Bearing Fault Diagnosis Method for Improved Prototype Networks [J]. Vibration and Shock, 2023, 42(20):214-221.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]//Advances in Neural Information Processing Systems. 2017, 30: 5998-6008.

[20] Zhang X, Li L, Shi H, et al. An ensemble Swin-LE model with residuals for rolling bearing fault diagnosis[J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024, 46(4): 211.

[21] Yan J, Zhu X, Wang X, et al. A new fault diagnosis method for rolling bearings based on the Swin Transformer and generalized s-Transform[J]. Mathematics, 2024, 13(1): 45-45.

[22] Xu P, Zhang L. A fault diagnosis method for rolling bearings based on the 1D-ViT model[J]. IEEE Access, 2023, 11: 39664–39674.

[23] Yang K, Chen X, Zhang L, et al. Cross-domain Bearing Fault Diagnosis Based on Joint Subdomain Comparison and Alignment[J]. China Mechanical Engineering, 2025, 36(5): 1065-1073.

[24] Bie F, Zhou Z, Li Q, et al. Research on a Rolling Bearing Fault Diagnosis Method Based on COA-CNN[J]. Noise and Vibration Control, 2025, 45(4): 136-142.

[25] Li Y, Yang J, Gao T, et al. Fault Diagnosis Method for Rotating Machinery Based on a Personalized Federal Domain Generalization Framework[J]. Journal of Instrumentation, 2025, 46(8): 75-86.

[26] Zhang J, Ding C, Jiang X, et al. Encoder Signal Feature Extraction Based on a Weighted Dual-Domain Sparse Decomposition Model and Its Application in Planetary Gearbox Fault Diagnosis[J]. Journal of Instrumentation, 2025, 46(7): 260-270.

[27] Yang Z, Cen J, Liu X, et al. Research on a Bearing Fault Diagnosis Method Based on a Transformer Neural Network [J]. Measurement Science and Technology, 2022, 33(8): 085111.

[28] Li A, Yao D, Yang J, et al. Bearing diagnosis using an anti-noise neural network based on selectable branch multiscale modules and attention mechanisms[J]. IEEE Sensors Journal, 2023, 24(5): 5830–5840.

[29] Chen H, Wei J, Huang H, et al. A novel imbalanced fault diagnosis method based on generative adversarial networks with balancing serial CNN and Transformer (BCTGAN)[J]. Expert Systems with Applications, 2024, 258: 125171.

[30] Fang H, Deng J, Chen D, et al. You can get smaller: A lightweight self-activation convolution unit modified by a Transformer for fault diagnosis[J]. Advanced Engineering Informatics, 2023, 55: 101890.

[31] Xu Z, Xuan Y, Wang H, Wang R. Fault Diagnosis of Rolling Bearings Based on VMD-CNN-BiTCN[J]. Small & Special Electrical Machines, 2025, 53(2): 68-73.

[32] Xu K, Liu D, Zong Z, Wang C, Wang R. Rolling Bearing Fault Diagnosis Based on FFT-CNN-Transformer[J]. Journal of Naval Aviation University, 2025, 40(6): 851-860.